

Context-Aware Mixture of Domain Experts for Bodily Expression of Emotion in the Wild

Mohammad Mahdi Dehshibi  Senior Member, IEEE and David Masip  Senior Member, IEEE

Abstract—The same body posture can convey entirely different emotions depending on its surrounding context, yet most methods for recognising bodily emotions treat scene and object cues as auxiliary feature augmentations rather than as structured priors over the plausibility of emotions. We introduce the Context-Aware Mixture of Domain Experts (CA-MoDE) for bodily emotion recognition. CA-MoDE incorporates dedicated scene and object experts to generate soft distributions over emotion categories conditioned on their respective domains. These domain-conditioned soft predictions serve as structured contextual priors that modulate the body expert’s predictions at the distributional level rather than at the feature level. To fuse these multi-domain signals, we propose a task-tailored max-endorsement gating strategy that selects the strongest contextual signal across experts for each emotion dimension. Our gating strategy mitigates the signal dilution that typically occurs when conflicting or uninformative context distributions are averaged. CA-MoDE achieves an Emotion Recognition Score of 0.3269 on the Body Language Database. By outperforming existing temporal models using only single still images, our framework demonstrates that explicitly modelling structured spatial context can serve as a complementary discriminative proxy for the behavioural dynamics typically captured by video.

Index Terms—Bodily Emotion Recognition, Context-aware Learning, Max-Endorsement Gating, Mixture of Domain Experts, Co-occurrence Prior.

1 INTRODUCTION

EMOTION is not a property of the body in isolation. The same posture can signal excitement or fear, aversion or affection, depending on the surrounding scene and objects. Psychological studies confirm that body language is systematically ambiguous without its surrounding context [1], and that the environment in which a gesture occurs can alter its emotional interpretation [2, 3, 4, 9, 11, 18]. Recognising bodily expressions of emotion in the wild, therefore, requires reasoning about context, not just about the body. Figure 1 illustrates this dependence of bodily emotion on scene and object context.

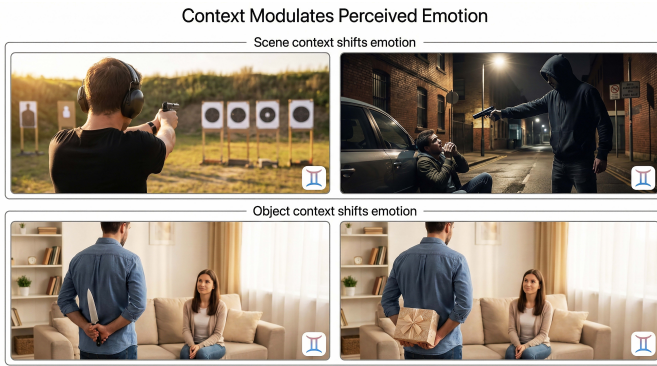


Fig. 1. Context fundamentally alters the perceived emotion of an action. The same object (a handgun) evokes pleasure and excitement in a shooting range, but fear and violence on a street. The same scene (a living room) evokes threat or affection depending on whether a partner holds a knife or a gift box. *These synthetic images are used for illustrative purposes only and do not depict real individuals or events.*

Department of Computer Science, University of the West of England, Bristol, U.K. Email: mohammad.dehshibi@uwe.ac.uk

Department of Computer Science, Universitat Oberta de Catalunya, Barcelona, Catalonia, Spain. Email: dmasipr@uoc.edu

Automatic Identification of Bodily Expression of Emotions (AIBEE) has advanced substantially since the introduction of large-scale benchmarks such as EMOTIC [12] and BoLD [15]. However, this progress has largely improved representation learning rather than changing how contextual evidence is modelled. In most architectures, contextual cues (e.g., scene categories and detected objects) are appended as auxiliary feature channels to a body-based backbone [6, 19]. The network is then expected to discover that, for instance, a detected weapon shifts the probability distribution from pleasure to fear. This design has a predictable failure mode. When scene context provides a strong signal and object context is silent, or vice versa, averaging or concatenation dilutes the informative cue with uninformative noise. When contextual labels are absent at test time, the network has no mechanism for reasoning about what the missing context implies.

To address this limitation, we propose CA-MoDE¹, a context-aware mixture of domain experts that treats context integration as probabilistic evidence rather than feature augmentation. CA-MoDE uses dedicated scene and object experts to predict soft pseudo-labels conditioned on their respective domains. These predictions are compiled into emotion-context co-occurrence priors over the training set. These priors encode both the emotions supported by available contextual cues and the information carried by unavailable or weak contextual evidence. We also introduce a max-endorsement gating strategy that uses these priors to modulate the output of an independent emotion expert. By selecting the strongest contextual signal for each emotion dimension, the gate mitigates the dilution that occurs when informative and uninformative cues are averaged.

CA-MoDE is fully differentiable and trained end-to-end on the Body Language Database (BoLD) [15]. It achieves an Emotion

1. To facilitate reproducibility, the PyTorch implementation of CA-MoDE is publicly available at <https://github.com/dehshibi/CA-MoDE>.

Recognition Score (ERS) of 0.3269 using only single still images. This gain does not imply superiority of static over dynamic analysis. It does suggest that when contextual cues are explicitly modelled as structured probabilistic priors, they can partly compensate for the absence of temporal information, especially for categorical emotion recognition.

2 RELATED WORK

Computational emotion recognition has historically been anchored in facial analysis. Early architectures exploited the Facial Action Coding System to define a tractable label space and achieved strong performance under controlled, frontal conditions [17]. Multi-task formulations later incorporated action unit detection to reduce the ambiguity of label-scarce settings [20]. These methods, however, failed to generalise in the wild. Psychological evidence shows that body language provides robust cues for emotion recognition in the wild, especially when faces are occluded or ambiguous [1].

Interest in bodily expressions of emotion grew from these observations. The body carries broad postural and kinematic signals [2], but their interpretation depends strongly on the surrounding scene and the objects involved. Large-scale databases such as BoLD [15] made bodily expressions of emotion a tractable computational task. Luo et al. [15] introduced BoLD and a multi-task framework that jointly addresses pose estimation and emotion recognition, and the model must implicitly learn the influence of context on emotion.

Subsequent work on BoLD expanded the context integration pipeline while preserving this structure. Filntis et al. [6] combined skeleton sequences with multi-segment body features in a late-fusion ensemble. Pikoulis et al. [19] incorporated scene attributes and optical flow over skeleton coordinates using a Spatial-Temporal Graph Convolutional Network. Li et al. [13] proposed SIB-Net, a hierarchical architecture that maps body crops against multi-scale background contexts through sequential interactive layers. A shared constraint of these models is also that contextual cues are not explicitly modelled, with their influence on the output distribution learned implicitly.

Recent advancements in Vision-Language Models (VLMs) have introduced an alternative paradigm for emotion recognition. Zhang et al. [24] applied contrastive learning on uncured video captions with subject-aware attention masking to train Emotion-CLIP, achieving a mean Average Precision (mAP) of 42.23% and a mean ROC Area (mRA) of 81.36% on BoLD. Similarly, Xenos et al. [22] demonstrated the effectiveness of semantic priors by using a pretrained VLM to generate scene descriptions, reporting an mAP of 26.66% and a classification accuracy of 93.08%. The improvement demonstrates that structured semantic priors over scene content are more informative than raw feature concatenation. These results are promising, but they leave two issues unresolved.

First, these VLM-based methods focus on specific classification subtasks of the BoLD benchmark rather than the full suite of continuous and discrete emotion predictions. Therefore, the unavailability of mRA prevents computing the composite ERS, and, subsequently, a direct quantitative comparison is not applicable. Moreover, these methods depend on VLMs and provide no explicit mechanism for reasoning about what unavailable or weak contextual evidence implies.

Second, beyond the issue of metric comparability, the remaining gap is structural. Across conventional and recent VLM-based approaches, context is used to enrich inputs or align representations, yet the probabilistic link between context and emotion is rarely made explicit. When one contextual source provides a strong signal, and another is silent, feature alignment or fusion can attenuate the informative cue. When contextual evidence is missing or weak at test time, there is no explicit mechanism for reasoning about what that absence implies. CA-MoDE addresses this gap by computing emotion-context co-occurrence statistics from training data and using them as conditional distributions in the gating mechanism.

3 PROPOSED METHOD

We formulate bodily emotion recognition as a multi-output regression problem. Given a dataset of n images $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, with $x_i \in \mathbb{R}^{\text{width} \times \text{height} \times 3}$ and $y_i \in \mathbb{R}^{d_y}$, $d_y = 29$, our goal is to train a deep neural network to predict the conditional expectation of the emotion labels given the input image. The target vector y_i comprises 26 discrete emotions mapped to $[0, 1]$ and three continuous affective variables (Valence, Arousal, Dominance; VAD) in $[1, 10]$. Figure 2(a) shows emotional category proportion and VAD measurement for a scene. We optimise the network end-to-end by minimising the mean squared error between the predictions and the ground-truth annotations.

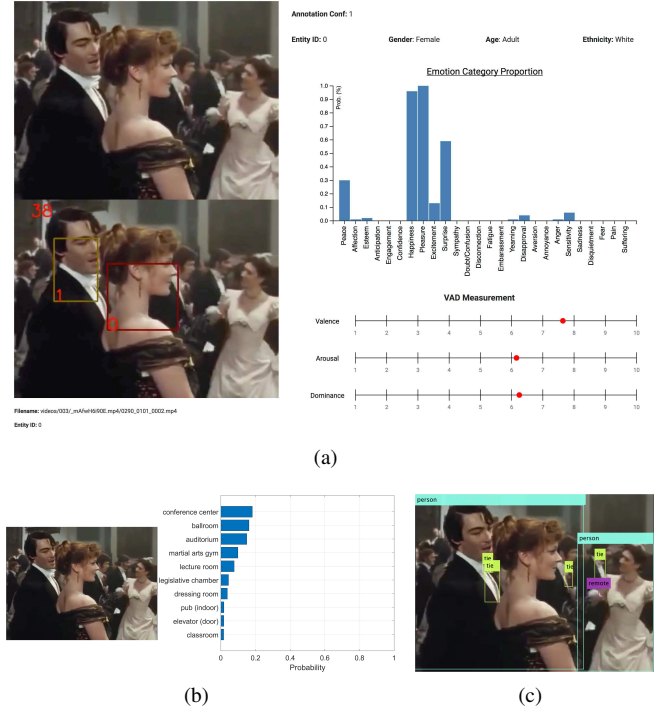


Fig. 2. (a) A sample from the BoLD database that mainly represents happiness. (b) Top-10 scene soft pseudo-labels obtained by applying Places-CNN, pre-trained on the Places2 database, to the input image. (c) Object soft pseudo-labels obtained by applying YOLO, pre-trained on the Microsoft COCO database, to the input image.

We assume that each image can be characterised by three complementary components, namely **emotion**, **place**, and **object**, each of which is handled by a dedicated domain expert. While emotion labels are explicitly provided in $\mathcal{D}$, contextual information (i.e., place and object) is not explicitly available in BoLD. To address

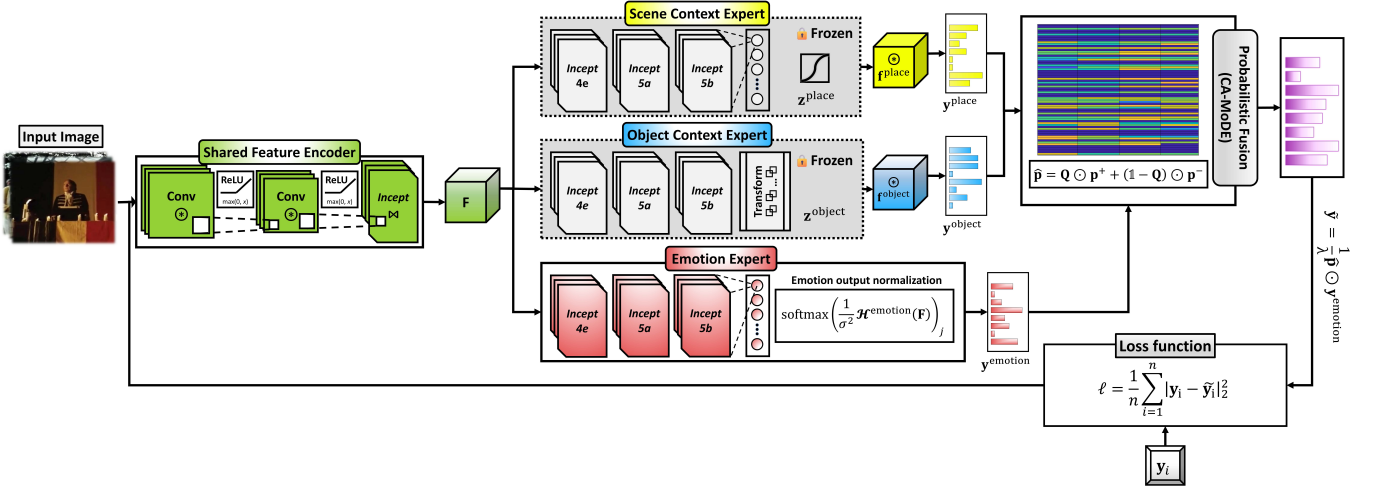


Fig. 3. The proposed CA-MoDE architecture. The shared feature encoder is trainable, while the scene and object context expert backbones are frozen after initialisation from pre-trained models. During training, gradients flow only through the Emotion Expert and the learnable 1×1 convolution filters that project the contextual outputs into the fusion space.

this, we generate **soft pseudo-labels** for scene and object context (see Figs. 2(b) and 2(c)) using domain experts pre-trained on the Places2 [25] and Microsoft COCO [14] databases, respectively.

Mensink et al. [16] showed that attribute co-occurrences observed during training tend to reappear at test time with high probability. Incorporating such co-occurrences helps reduce unwanted outliers introduced by domain integration and supports the use of L_2 as the regression loss for this task [23]. However, directly fine-tuning the contextual experts on BoLD is not possible because place and object labels are unavailable.

To address this problem, we propose a Context-Aware Mixture of Domain Experts (CA-MoDE) in which the probabilistic fusion mechanism channels contextual information into the emotion learning process using *a priori* knowledge about the joint probability of emotions and both available and anticipated unavailable places and objects. We formulate a differentiable probabilistic fusion scheme, grounded in conditional probability identities, to combine the uncertain contextual information with the predicted image-based emotional states. Our fusion allows the network to discover correlations between context and emotion without additional post-processing or domain-specific regularisation. Figure 3 illustrates the proposed architecture.

3.1 Context-Aware Mixture of Domain Experts Architecture

The proposed CA-MoDE comprises a shared feature encoder and three domain experts that process the extracted features in parallel. As illustrated in Fig. 3, the shared encoder provides a common feature representation to the scene context expert, the object context expert, and the emotion expert. The outputs of the contextual experts serve as soft pseudo-labels that feed into the probabilistic fusion mechanism described in Section 3.2.

(1) Shared Feature Encoder: All domain experts share a common feature encoder $\mathcal{H}^{\text{base}}$ based on GoogLeNet [21], where the output of the Inception 4d module serves as a shared feature representation. More formally, the shared feature map is computed as in Eq. (1).

$$\mathbf{F} = \mathcal{H}^{\text{base}}(\mathbf{X}), \quad \mathbf{F} \in \mathbb{R}^{W \times H \times D}, \quad (1)$$

where $W = 14$, $H = 14$, and $D = 528$ denote the width, height, and number of channels of the feature map, respectively. The tensor $\mathbf{F}$ is then fed to three domain experts specialised in scene context, object context, and bodily emotion cues.

(2) Scene Context Expert: The scene context expert $\mathcal{H}^{\text{place}}$ is responsible for describing the environment in which the action takes place. It resembles the architecture of Places-CNN with a GoogLeNet backbone and includes the Inception 4e, 5a, and 5b modules, followed by global average pooling, dropout, a fully connected layer, and a softmax classifier. The scene expert produces a probability distribution over 365 scene categories as formulated in Eq. (2).

$$\mathbf{z}^{\text{place}} = \mathcal{H}^{\text{place}}(\mathbf{F}), \quad \mathbf{z}^{\text{place}} \in \mathbb{R}^{1 \times 1 \times 365}, \quad (2)$$

where $\mathbf{z}^{\text{place}}$ contains soft pseudo-labels for scene context, i.e., predicted probabilities over Places2 categories. To harmonise the dimensionality of the contextual representations and prepare them for the probabilistic fusion described in Section 3.2, we apply a learnable 1×1 convolution followed by a bias term, as defined in Eq. (3).

$$\mathbf{y}^{\text{place}} = \mathbf{z}^{\text{place}} \otimes \mathbf{f}^{\text{place}} + \mathbf{b}^{\text{place}}, \quad (3)$$

where $\otimes$ denotes convolution, $\mathbf{f}^{\text{place}} \in \mathbb{R}^{1 \times 1 \times 365 \times \kappa}$ is a trainable filter bank, $\mathbf{b}^{\text{place}} \in \mathbb{R}^{\kappa}$ is a bias term, and κ is the chosen latent dimensionality for contextual features. The weights of $\mathcal{H}^{\text{place}}$ are initialized from a Places2 pre-trained model and kept frozen during training on BoLD; only $\mathbf{f}^{\text{place}}$ and $\mathbf{b}^{\text{place}}$ are updated.

(3) Object Context Expert: The object context expert $\mathcal{H}^{\text{object}}$ captures the set of objects present in the scene that may influence the perceived emotion. This expert is inspired by the YOLO detection architecture and consists of three groups of convolutional, ReLU, and batch normalisation layers. In the first group (Inception 4e), the filter size is set to 14×14 to match the spatial resolution

of $\mathbf{F}$, whereas the second and third groups (Inception 5a and 5b) use 7×7 filters to facilitate the detection of smaller objects. The object expert produces a probability distribution over 80 object categories as given in Eq. (4).

$$\mathbf{z}^{\text{object}} = \mathcal{H}^{\text{object}}(\mathbf{F}), \quad \mathbf{z}^{\text{object}} \in \mathbb{R}^{1 \times 1 \times 80}, \quad (4)$$

where $\mathbf{z}^{\text{object}}$ contains soft pseudo-labels for object context derived from a COCO pre-trained model. As in the scene expert, we project these probabilities into a κ -dimensional latent space using a 1×1 convolution, as defined in Eq. (5).

$$\mathbf{y}^{\text{object}} = \mathbf{z}^{\text{object}} \oplus \mathbf{f}^{\text{object}} + \mathbf{b}^{\text{object}}, \quad (5)$$

with $\mathbf{f}^{\text{object}} \in \mathbb{R}^{1 \times 1 \times 80 \times \kappa}$ and $\mathbf{b}^{\text{object}} \in \mathbb{R}^{\kappa}$. The parameters of $\mathcal{H}^{\text{object}}$ are initialised from a COCO pre-trained detector and kept frozen during training on BoLD; only $\mathbf{f}^{\text{object}}$ and $\mathbf{b}^{\text{object}}$ are updated. The vectors $\mathbf{y}^{\text{place}}$ and $\mathbf{y}^{\text{object}}$ constitute compact contextual representations that will be used in Section 3.2 to estimate emotion-context co-occurrences and drive the probabilistic fusion mechanism.

(4) Emotion Expert: The emotion expert $\mathcal{H}^{\text{emotion}}$ learns the mapping between the shared feature map $\mathbf{F}$ and the 29-dimensional emotion vector. It reuses the Inception 4e, 5a, and 5b modules, followed by global average pooling, dropout, and a fully connected layer. The output is formulated in Eq. (6).

$$\mathbf{y}^{\text{emotion}} = \mathcal{H}^{\text{emotion}}(\mathbf{F}), \quad \mathbf{y}^{\text{emotion}} \in \mathbb{R}^{1 \times d_y}. \quad (6)$$

Although emotions can co-occur, we deliberately apply a precision-weighted softmax over the logits to sharpen the distribution toward the dominant emotion. This mechanism produces a discriminative prior for the gating module. Let $H_j(\mathbf{F})$ denote the j -th logit. The scaled class probabilities are given by Eq. (7).

$$p(y = j \mid \mathcal{H}^{\text{emotion}}(\mathbf{F}), \sigma) = \text{softmax}\left(\frac{1}{\sigma^2} \mathcal{H}^{\text{emotion}}(\mathbf{F})\right)_j, \quad (7)$$

where $\sigma > 0$ is a learnable scale parameter. Because the logits are multiplied by $1/\sigma^2$, a smaller σ produces a sharper, more peaked distribution, while a larger σ produces a softer distribution. The parameters of $\mathcal{H}^{\text{emotion}}$ are updated during training on BoLD using the multi-output regression loss defined in Section 4, unlike the contextual experts whose backbones are frozen.

3.2 Probabilistic Fusion of Contextual Experts

We now derive the probabilistic fusion mechanism that modulates emotion predictions using contextual cues. To streamline notation, let $\mathbb{B}$ denote the emotion variable and $\mathbb{A}_1, \mathbb{A}_2$ denote scene and object contexts, respectively. These correspond to outputs from the emotion expert ($\mathcal{H}^{\text{emotion}}$), scene expert ($\mathcal{H}^{\text{place}}$), and object expert ($\mathcal{H}^{\text{object}}$).

Estimating Emotion-Context Co-occurrence: We first compute empirical co-occurrence statistics between emotions and contextual cues over the training set. These statistics serve as priors for our probabilistic fusion mechanism. Let N_i be the accumulated soft counts of the i -th emotion label over the training set, and n the total number of training samples. We define the empirical prior probability of the i -th emotion by Eq. (8).

$$p_i = \Pr(\mathbb{B}_i) = \frac{N_i}{n}, \quad 1 \leq i \leq d_y. \quad (8)$$

Similarly, let $N_{i,j}$ denote the weighted accumulated soft count between the i -th emotion and the j -th contextual expert, where $j \in \{1, 2\}$ indexes the scene and object experts, respectively. For each sample, we summarise the output of expert j by a scalar confidence score $s_{k,j} \in [0, 1]$. This score, defined as the maximum predicted category probability, is then used to compute $N_{i,j} = \sum_{k=1}^n y_{k,i} \times s_{k,j}$. The corresponding joint probability estimate is calculated using Eq. (9).

$$\mathcal{C}_{i,j} = \Pr(\mathbb{B}_i \cap \mathbb{A}_j) = \frac{N_{i,j}}{n}. \quad (9)$$

Using these priors, we can approximate the conditional probability of an emotion given a contextual event using Eq. (10).

$$\mathcal{P}_{i,j}^+ = \Pr(\mathbb{B}_i \mid \mathbb{A}_j) = \frac{\Pr(\mathbb{B}_i \cap \mathbb{A}_j)}{\Pr(\mathbb{A}_j)} = \frac{\mathcal{C}_{i,j}}{q_j}, \quad (10)$$

where $q_j = \Pr(\mathbb{A}_j)$ is the prior probability of the j -th contextual event. It is estimated from training data by aggregating the predicted soft pseudo-labels over the dataset. We collect these conditional probabilities into $\mathcal{P}^+ \in \mathbb{R}^{d_y \times 2}$ as in Eq. (11).

$$\mathcal{P}^+ = \begin{bmatrix} \Pr(\mathbb{B}_1 \mid \mathbb{A}_1) & \Pr(\mathbb{B}_2 \mid \mathbb{A}_1) & \cdots & \Pr(\mathbb{B}_{d_y} \mid \mathbb{A}_1) \\ \Pr(\mathbb{B}_1 \mid \mathbb{A}_2) & \Pr(\mathbb{B}_2 \mid \mathbb{A}_2) & \cdots & \Pr(\mathbb{B}_{d_y} \mid \mathbb{A}_2) \end{bmatrix}^\top, \quad (11)$$

where the first and second columns correspond to scene and object context, respectively.

Modelling Anticipated Unavailable Context: While $\mathcal{P}^+$ captures what available context makes likely, the absence or low confidence of a contextual cue can also be informative. We refer to this signal as anticipated unavailable context and model it as $\mathcal{P}_{i,j}^- = \Pr(\mathbb{B}_i \mid \neg \mathbb{A}_j)$. Using standard conditional probability theory, we express $\Pr(\mathbb{B}_i \mid \neg \mathbb{A}_j)$ in terms of marginal and joint probabilities using Eq. (12).

$$\Pr(\mathbb{B}_i \mid \neg \mathbb{A}_j) = \frac{\Pr(\mathbb{B}_i \cap \neg \mathbb{A}_j)}{\Pr(\neg \mathbb{A}_j)} = \frac{\Pr(\mathbb{B}_i) - \Pr(\mathbb{B}_i \cap \mathbb{A}_j)}{1 - q_j}. \quad (12)$$

Substituting our empirical estimates leads to $\mathcal{P}_{i,j}^- = \frac{p_i - \mathcal{C}_{i,j}}{1 - q_j}$, where we collect these values into a matrix $\mathcal{P}^- \in \mathbb{R}^{d_y \times 2}$. This formulation allows us to reason not only about *what is present* in the scene, but also about *what is expected to be absent*, both of which can significantly influence perceived emotions.

Probabilistic Fusion of Contextual Priors: Given $\mathcal{P}^+, \mathcal{P}^-$, we first aggregate the contextual contributions for each emotion using Eq. (13).

$$p_i^+ = \max_{j \in \{1,2\}} \mathcal{P}_{i,j}^+, \quad p_i^- = \max_{j \in \{1,2\}} \mathcal{P}_{i,j}^-, \quad (13)$$

and stack these into vectors $\mathbf{p}^+, \mathbf{p}^- \in \mathbb{R}^{1 \times d_y}$. The max operation implements a winner-take-all selection for each emotion dimension, mitigating dilution of strong single-expert signals. These aggregated vectors are then combined by a gate that balances available contextual support against the missing-context prior.

Gating Weights from Contextual Support: To combine $\mathbf{p}^+$ and $\mathbf{p}^-$, we define a gating vector $\mathbf{Q} \in [0, 1]^{1 \times d_y}$ that, for each emotion dimension i , determines how much weight to assign to the available-context prior versus the missing-context prior. Let Q_i denote the i -th entry of $\mathbf{Q}$. Intuitively, if the aggregated available endorsement p_i^+ is strong, then Q_i should be high, giving more weight to available context. Conversely, if p_i^+ is weak, Q_i should be low, and the probabilistic fusion should rely more on the anticipated unavailable context p_i^- . We define Q_i as in Eq. (14).

$$Q_i = \text{sigmoid}(\alpha(p_i^+ - \tau)), \quad 1 \leq i \leq d_y. \quad (14)$$

The sigmoid gate maps the strongest available contextual endorsement per emotion, p_i^+ , to a soft weight in $(0, 1)$, where τ controls the threshold above which contextual support is considered strong and α controls the sharpness of the transition. We fix $\tau = 0.5$ as the threshold and $\alpha = 10$ to control gate sharpness across all experiments. These values were selected via validation set grid search over $\tau \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and $\alpha \in \{5, 10, 15, 20\}$.

Under this gate, higher Q_i values indicate stronger contextual endorsement and make the fused estimate rely more on the available-context term p_i^+ , whereas lower values shift it towards the anticipated unavailable-context term p_i^- . Empirically, this simple interpolation provides a favourable balance between simplicity and robustness, especially in the presence of noisy soft pseudo-labels. We then compute a convex combination of the two contextual priors for each emotion using Eq. (15).

$$\hat{p}_i = Q_i p_i^+ + (1 - Q_i) p_i^-, \quad 1 \leq i \leq d_y. \quad (15)$$

In vector form, $\hat{\mathbf{p}} = \mathbf{Q} \odot \mathbf{p}^+ + (1 - \mathbf{Q}) \odot \mathbf{p}^-$, where $\odot$ denotes element-wise multiplication and $\mathbf{1}$ is the all-ones vector of size $1 \times d_y$. The fused vector $\hat{\mathbf{p}}$ encodes, for each emotion, the combined influence of available and missing contextual evidence, weighted by the strength of contextual endorsement. It is grounded in conditional probability identities but designed as a robust, differentiable heuristic rather than a strict Bayesian posterior.

Emotion Prediction and Training Objective: Finally, the fused contextual vector $\hat{\mathbf{p}}$ is used to modulate the output of the emotion expert, reinforcing emotion predictions that are consistent with the contextual priors while suppressing those that are not. The predicted emotion vector is defined using Eq. (16).

$$\tilde{\mathbf{y}} = \frac{1}{\lambda} \hat{\mathbf{p}} \odot \mathbf{y}^{\text{emotion}}, \quad (16)$$

where $\lambda > 0$ is a scalar hyperparameter selected on the validation set to control the scale of the predictions. Intuitively, entries of $\hat{\mathbf{p}}$ close to one amplify the corresponding emotion predictions, while entries close to zero suppress predictions that neither the available nor the anticipated unavailable context supports.

Given ground-truth labels $\{\mathbf{y}_i\}_{i=1}^n$ from BoLD, we train the complete network end-to-end by minimizing the mean squared error using Eq. (17).

$$\ell = \frac{1}{n} \sum_{i=1}^n \|\mathbf{y}_i - \tilde{\mathbf{y}}_i\|_2^2. \quad (17)$$

Following [15], we obtain classification metrics (mAP, mRA) by thresholding $\tilde{\mathbf{y}}_{1:26}$ at 0.5. We directly evaluate the regression

metric (mR²) on $\tilde{\mathbf{y}}_{27:29}$. The gradients propagate through the probabilistic fusion mechanism and the emotion expert, allowing the network to adapt the mapping from shared features and contextual signals to the final emotion representation.

4 ARCHITECTURE DETAILS

Experiments were performed on the NVIDIA GeForce RTX 5090 GPU (24 GB). All models were implemented in PyTorch and trained from scratch or fine-tuned as described below. All hyperparameter selection is performed on the validation set, with final evaluation on the held-out test set.

Scene Context Expert: In $\mathcal{H}^{\text{place}}$, we initialise the weights and parameters of all layers to those of Places-CNN, which was pre-trained on the Places2 database [25]. We set the learning rates of all its layers to zero to ensure that the output distribution of $\mathcal{H}^{\text{place}}$ remains consistent with the scene soft pseudo-labels used to compute $\mathcal{P}^+$ and $\mathcal{P}^-$ during training.

Object Context Expert: In $\mathcal{H}^{\text{object}}$, we replace the pooling layers with strided convolutions of strides 16×16 and 32×32 , respectively, to preserve spatial information while reducing the feature map resolution. We pre-train $\mathcal{H}^{\text{base}} + \mathcal{H}^{\text{object}}$ jointly on the Microsoft COCO database [14] using a YOLO-style configuration with 7 anchor boxes and a non-maximal suppression threshold of 0.4. We then transfer the weights of $\mathcal{H}^{\text{object}}$ to CA-MoDE and set the learning rates to zero, ensuring the output distribution remains consistent with the object soft pseudo-labels used to compute $\mathcal{P}^+$ and $\mathcal{P}^-$.

Emotion Expert: In $\mathcal{H}^{\text{emotion}}$, we train the network using stochastic gradient descent with a momentum of 0.9. The initial learning rate is set to 10^{-2} and is decreased by a factor of 0.1 every 45 epochs, with a maximum of 90 training epochs. We use a mini-batch of 8 per iteration, and the training data is shuffled before each epoch. The training parameters for the shared feature encoder $\mathcal{H}^{\text{base}}$ are set to the same values.

Hyperparameter selection: The CA-MoDE architecture has two tuned hyperparameters, κ and λ . The parameter κ controls the output dimensionality of the learnable filter banks in the context experts. The parameter λ controls the scale of the final prediction $\tilde{\mathbf{y}}$. To determine the appropriate range for κ , we apply the pre-trained scene and object context experts to the training set.

The context experts map input images x_i to their respective class. For the scene context expert, the softmax output, z_i^{place} , is a 365-dimensional vector containing the probability of each scene category. For the object context expert, the softmax output, z_i^{object} , has 80 entries, one per COCO category. The list of detected objects per image may contain multiple instances of the same category with different confidence scores. To produce a fixed-length representation, we assign the normalised cumulative confidence score of all detected instances to the corresponding entry and set the remaining entries to zero.

The maximum value of κ is bounded by $\min(365, 80) = 80$. To identify the minimum meaningful value of κ , we compute the normalised average soft pseudo-label vectors over the training set for both scene and object contexts, whose distributions are shown in Figs. 4(b) and 4(c), respectively. Applying a threshold of 0.01, we find that 27 place and 14 object categories exceed this

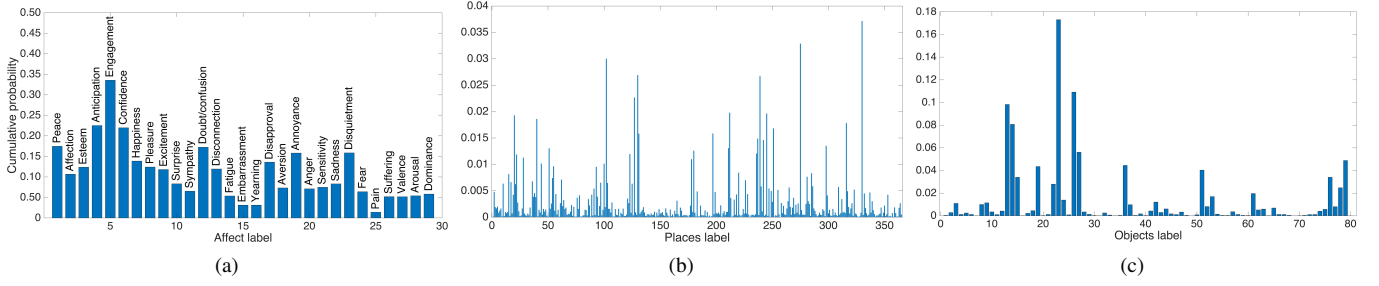


Fig. 4. Cumulative probability of (a) emotion labels in the BoLD database, (b) scene soft pseudo-labels obtained by applying the scene context expert to the BoLD training set, and (c) object soft pseudo-labels obtained by applying the object context expert to the BoLD training set.

threshold, demonstrating that the vast majority of entries in both vectors are negligible. We therefore evaluate κ from 14 to 80 in steps of 6.

Similarly, we evaluate the impact of λ by varying it from 0 to 0.5 in steps of 0.1. The analysis of the trade-offs across all metrics on the validation set shows that $(\kappa, \lambda) = (56, 0.2)$ allows CA-MoDE to achieve an ERS of 0.4122 (see Fig. 5). We use these values in all subsequent experiments.

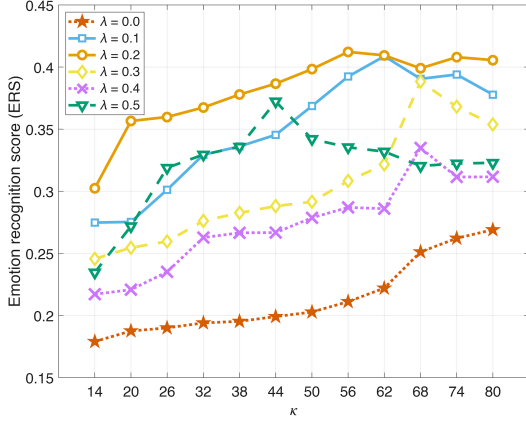


Fig. 5. Emotion Recognition Score (ERS) on the validation set for CA-MoDE as a function of (κ, λ) . The best trade-off is achieved at $(\kappa, \lambda) = (56, 0.2)$ with an ERS of 0.4122.

From Fig. 5 and an analysis of prediction errors, we observe that larger values of (κ, λ) not only increase computational complexity but also make the prediction $\tilde{y}$ more sensitive to the contextual modulation terms. This increased sensitivity reduces robustness when contextual soft pseudo-labels are noisy or uninformative.

5 EXPERIMENTS

5.1 Body Language Database (BoLD)

Body Language Dataset (BoLD) [15] is constructed from video clips drawn from the AVA database [7], a collection of YouTube movie clips comprising 9,876 clips and 13,239 isolated human instances. BoLD annotations^{2,3} were obtained through crowd-

2. The discrete categorical labels include Peace, Affection, Esteem, Anticipation, Engagement, Confidence, Happiness, Pleasure, Excitement, Surprise, Sympathy, Doubt/Confusion, Disconnection, Fatigue, Embarrassment, Yearning, Disapproval, Aversion, Annoyance, Anger, Sensitivity, Sadness, Disquietment, Fear, Pain, and Suffering.

3. The continuous affective variables (VAD) are Valence (V), Arousal (A), and Dominance (D).

sourcing on Amazon Mechanical Turk to establish multi-annotator consensus for highly subjective emotional states.

5.2 Evaluation Metrics and Experimental Protocol

We use the mean R^2 score to evaluate the regression component of the proposed model. For the k -th continuous emotion dimension, R^2 measures the ratio of variance explained by the model to the total variance of the ground truth, as defined in Eq. (18).

$$R^2(y^{(k)}, \tilde{y}^{(k)}) = 1 - \frac{\sum_{i=1}^n (y_i^{(k)} - \tilde{y}_i^{(k)})^2}{\sum_{i=1}^n (y_i^{(k)} - \varepsilon^{(k)})^2}, \quad \varepsilon^{(k)} = \frac{1}{n} \sum_{i=1}^n y_i^{(k)}, \quad (18)$$

where $y_i^{(k)}$ and $\tilde{y}_i^{(k)}$ are the ground-truth and predicted values of the i -th sample for dimension k , respectively. The mean coefficient of determination is then $mR^2 = \frac{1}{3} \sum_{k=1}^3 R^2(y^{(k)}, \tilde{y}^{(k)})$, averaged over the VAD dimensions.

Since we formulate the AIBEE task as a multi-output regression problem, classification metrics for the 26 discrete emotions are obtained by thresholding the continuous predictions $\tilde{y}_{1:26}$ at 0.5. For each of the 26 emotion dimensions ($k \in \{1, \dots, 26\}$), we compute Average Precision $AP^{(k)}$ and Receiver Operating Characteristic Area $RA^{(k)}$, and report their means. For ease of comparison across methods, the Emotion Recognition Score (ERS) aggregates regression and classification performance into a single scalar using Eq. (19).

$$\text{ERS} = \frac{1}{2} \left(mR^2 + \frac{1}{2} (mAP + mRA) \right). \quad (19)$$

We use the standard BoLD evaluation protocol, partitioning the dataset into training (60%), validation (10%), and test sets (30%). The partitioning was performed at the clip level to ensure that no two frames from the same video are assigned to different sets, while preserving the BoLD class distribution within each split. All hyperparameter selection is performed on the validation set, and the final evaluation is reported exclusively on the held-out test set.

5.3 Experimental results

Table 1 reports the performance of CA-MoDE alongside competitive methods on the BoLD test set. Following the evaluation protocol of [15], we include a random method based on priors (Chance) as a lower-bound reference. The comparison includes the BoLD baseline [15], the BEEU challenge winner [6], and the top-ranked method on the BoLD leaderboard [19], all of which

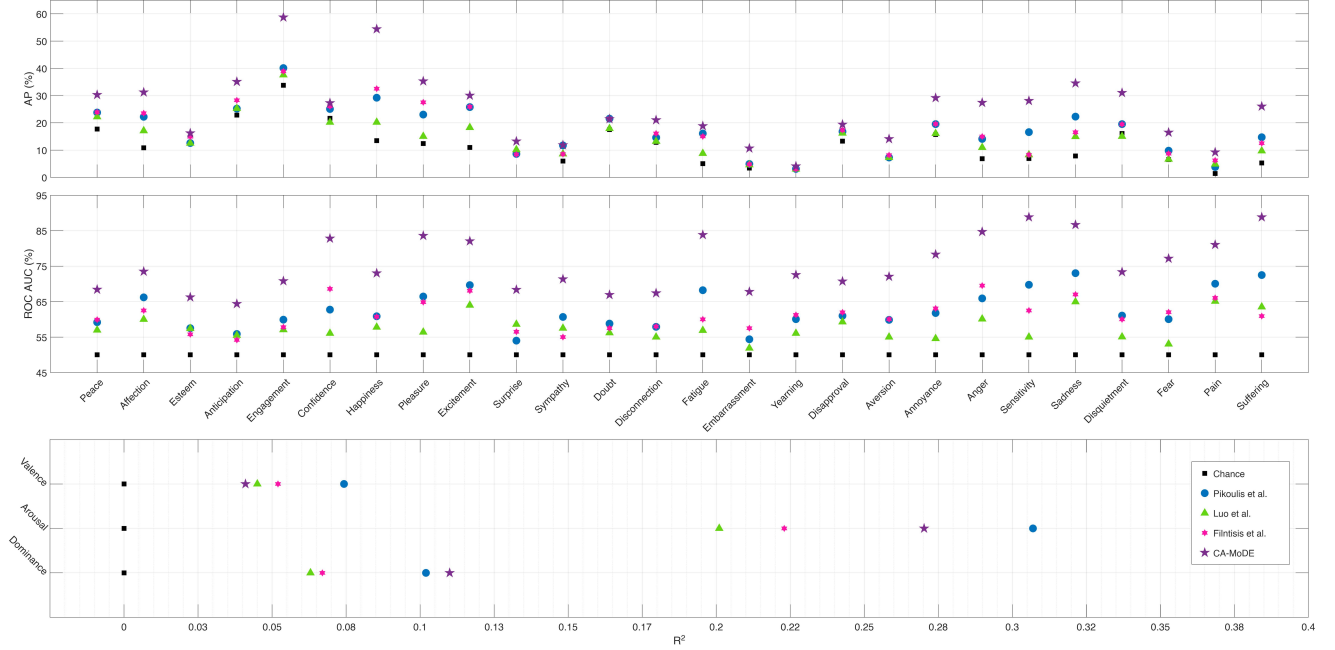


Fig. 6. Per-emotion metric comparisons on the BoLD test set. Classification performance for discrete emotions is reported based on the mAP in the first row and mRA in the second row. Regression performance for the VAD is reported based on the mR^2 score in the third row.

leverage temporal information across video frames through optical flow, skeleton sequences, or multi-segment sampling, and all of which report the three metrics required to compute ERS.

TABLE 1
Performance on the BoLD test set.

Method	ERS	Regression	Classification	
		$mR^2 \uparrow$	$mAP \uparrow$	$mRA \uparrow$
Chance	0.1513	0	0.1055	0.5000
Luo et al. [15]	0.2531	0.1030	0.1714	0.6352
Filntsis et al. [6]	0.2623	0.1141	0.1796	0.6416
Pikoulis et al. [19]	0.3058	0.1609	0.2187	0.6829
CA-MoDE	0.3269	0.1404	0.2938	0.7333

What makes this result noteworthy is that all three competing methods exploit temporal information across video frames using optical flow, multi-segment sampling, or skeleton sequences. In contrast, CA-MoDE operates on single still images and relies exclusively on scene and object context for its reasoning. The fact that a still-image method can match or even surpass video-based approaches on overall ERS and, by a clear margin, on classification metrics, highlights the contribution of spatial contextual cues relative to temporal dynamics.

The lower mR^2 reflects a deliberate architectural trade-off. CA-MoDE conditions emotion predictions on discrete contextual evidence from scene and object experts. This conditioning is more beneficial for classifying categorical emotions than for regressing continuous dimensions. The resulting gains in mAP and mRA over the previous best are sufficiently large that the aggregate ERS improves despite the regression deficit.

Figure 6 provides per-emotion comparisons across all methods. The emotion categories Engagement, Happiness, Pleasure, Anticipation, Sadness constitute the top-5 most accurately predicted emotions. This is particularly notable given that the BoLD database is biased towards Engagement, Anticipation, Confidence, Peace,

Doubt/Confusion (see Fig. 4(a)). The top-5 predicted categories only partially overlap with the five most frequent categories. This divergence suggests that the probabilistic fusion mechanism in CA-MoDE attenuates the effect of label imbalance by conditioning predictions on contextual evidence rather than on marginal label frequency alone. In the ablation study, we demonstrate that explicitly modelling both available and anticipated unavailable context directly contributes to this behaviour.

In the assessment of continuous emotions, all methods demonstrate stronger regression performance for arousal than for valence and dominance, consistent with the subjective human evaluation gap reported in [15]. For CA-MoDE specifically, this asymmetry reflects the unequal availability of static cues. Arousal correlates with postural and scene-level features that a single frame can convey, whereas valence and dominance require finer disambiguation that P^+ and P^- only partially provide. The modest mR^2 values across all methods confirm that the full VAD regression task remains an open challenge beyond the reach of spatial contextual reasoning alone.

5.4 Ablation study

In the ablation study, we conducted six sets of experiments to quantify the contribution of each component of CA-MoDE:

(1) **Contribution of Underlying Network Architecture:** To assess whether CA-MoDE’s performance depends on a specific network family, we replace the GoogLeNet-derived modules across the shared encoder and all three domain experts with ResNet-18 [8] and EVA-02 [5] equivalents. Table 2 reports the trainable parameter count for each configuration, together with ERS, regression, and classification performance. EVA-02 achieves marginally higher ERS and mAP , consistent with the benefits of large-scale transformer pre-training, but at roughly 45 times the trainable parameter count of the proposed architecture. The proposed architecture matches this performance

TABLE 2
Impact of the underlying network architecture on AIBEE performance.

Architecture Family	Trainable Parameters (M)	ERS	Regression	Classification	
			mR^2	mAP	mRA
Ablation					
ResNet-18 [8]	10.51	0.3206	0.1398	0.2815	0.7216
EVA-02 [5]	273.15	0.3272	0.1392	0.3067	0.7238
Proposed architecture					
CA-MoDE (GoogLeNet-derived)	6.11	0.3269	0.1404	0.2938	0.7333

while outperforming both alternatives on regression and mRA , using substantially fewer trainable parameters. These results indicate that CA-MoDE’s gains are attributable to the proposed probabilistic fusion mechanism rather than to the capacity of the underlying network architecture, and that the formulation does not require large-capacity networks to achieve competitive performance.

(2) Contribution of Pre-trained Model Weight: In this experiment, instead of assigning random weights to the $\mathcal{H}^{base} + \mathcal{H}^{emotion}$, we evaluated GoogLeNet weights pre-trained on ImageNet, Places2, and Microsoft COCO. We reduced the training schedule to 45 epochs while retaining all other parameters as described in Section 4. The results in Table 3 show that random initialisation leads to marginally better performance across all metrics, suggesting that BoLD’s visual statistics are sufficiently distinct from those of general-purpose databases.

TABLE 3
Impact of initial weights on AIBEE performance.

Initial weight	ERS	Regression	Classification	
		mR^2	mAP	mRA
Ablation				
Random	0.3041	0.1141	0.2901	0.6982
ImageNet	0.2990	0.1218	0.2889	0.6636
Places2	0.2956	0.1103	0.2702	0.6918
Microsoft COCO	0.2858	0.1097	0.2836	0.6402
Proposed architecture				
CA-MoDE	0.3269	0.1404	0.2938	0.7333

(3) Contribution of $\mathcal{H}^{place}$ and $\mathcal{H}^{object}$: Having confirmed that the results are not driven by initialisation, in this experiment, we analysed which of the two context experts contributes more. We independently ablated the scene context expert ($\mathcal{H} \ominus \mathcal{H}^{place}$) and the object context expert ($\mathcal{H} \ominus \mathcal{H}^{object}$) by removing each in turn and retraining on BoLD. All parameters were retained as described in Section 4, except that κ was set to 100 when only the scene context expert was active and to 40 when only the object context expert was active. The results are presented in Table 4.

TABLE 4
Ablation results for removing scene and object context experts on BoLD.

Architecture	ERS	Regression	Classification	
		m R^2	mAP	mRA
Ablation				
$\mathcal{H} \ominus \mathcal{H}^{\text{object}}$	0.3104	0.1286	0.2763	0.7081
$\mathcal{H} \ominus \mathcal{H}^{\text{place}}$	0.2993	0.1233	0.2606	0.6902
Proposed architecture				
CA-MoDE	0.3269	0.1404	0.2938	0.7333

Ablating the scene expert reduces ERS to 0.2993, while ablating the object expert reduces it to 0.3104. This gap suggests that scene context provides a stronger cue. We attribute the asymmetry to two factors. First, many BoLD clips come from older movies with lower image resolution, which inherently decreases the accuracy of $\mathcal{H}^{object}$ and propagates detection errors through the probabilistic fusion mechanism. Second, *person* dominates the detected categories in most clips, leaving most entries of $\mathbf{z}^{object}$ sparse after filtering and reducing the discriminative content of $\mathbf{y}^{object}$.

Even so, the object expert remains useful because neither expert alone approaches the full model’s ERS. The observed asymmetry also motivates our probabilistic fusion design. One expert typically dominates within a given emotion dimension, but the dominant expert can change across dimensions. For this reason, selecting the strongest contextual signal per dimension is preferable to averaging, which may dilute the leading signal with a weaker one.

(4) Contribution of Fusion Strategies: Given that both context experts contribute independently to recognition performance, we examine how their outputs should be integrated and how the gating signal should be computed. We compare the proposed probabilistic fusion to (i) an intermediate fusion strategy, in which the outputs of $\mathcal{H}^{place}$ and $\mathcal{H}^{object}$ are concatenated with $\mathcal{H}^{base}$ features and passed through a fully connected layer before being fed to $\mathcal{H}^{emotion}$; (ii) a late fusion variant that drops the anticipated unavailable context term; and (iii) the multi-task uncertainty weighting approach proposed by Kendall et al. [10].

Table 5 reports results for three fusion strategies. Intermediate fusion produces the weakest uncertainty profile because it exposes $\mathcal{H}^{emotion}$ directly to noise in the contextual outputs, a sensitivity amplified by the object detection sparsity reported in Table 4. A late fusion variant that drops the $\mathbf{p}^-$ (i.e., anticipated unavailable context) term improves confidence but lowers mAP compared to intermediate and late [10] fusion strategies. The positive co-occurrence term $\mathcal{P}^+$ reinforces contextually endorsed emotions with high confidence. However, when co-occurrence statistics are sparse or incorrect, those confident endorsements produce false positives that mAP , unlike the rank-based ERS, directly penalises. The suppression term $\mathcal{P}^-$ in CA-MoDE addresses this by attenuating implausible predictions. The Kendall et al. [10] variant reweights losses using learned task uncertainty rather than reshaping the output distribution. While the implicit loss regularisation improves mAP , the ERS remains below the probabilistic late fusion strategy.

(5) Contribution of Gating Mechanism: For gating, we substitute the proposed max-endorsement with the mean and learned variants. In all variants, the gate retains the same shifted-sigmoid

TABLE 5

Effect of fusion strategies on prediction confidence and agreement with ground truth. Lower E indicates more confident predictions, and higher MI indicates stronger dependence between predictions and ground truth.

Fusion method	ERS	Regression	Classification		Uncertainty	
		mR^2	mAP	mRA	E ↓	MI ↑
Ablation						
Intermediate fusion	0.3041	0.1226	0.2813	0.6902	1.66	3.70
Late fusion ($\hat{\mathbf{p}} = \mathbf{Q} \odot \mathbf{p}^+$)	0.3133	0.1358	0.2715	0.7103	1.39	4.28
Late fusion (Kendall et al. [10])	0.3103	0.1291	0.2828	0.7003	1.51	3.84
Proposed architecture						
CA-MoDE	0.3269	0.1404	0.2938	0.7333	1.25	4.51

TABLE 6

Impact of the gating mechanism. All variants use the same CA-MoDE architecture. The comparison isolates the gating mechanism by varying only how $\mathbf{Q}$ is computed across variants.

Gating variant	ERS	Regression	Classification		Uncertainty	
		mR ²	mAP	mRA	E ↓	MI ↑
Ablation						
Mean gating	0.2994	0.1249	0.2670	0.6809	1.53	3.86
Learned gating	0.3070	0.1303	0.2703	0.6974	1.37	4.01
Proposed architecture						
CA-MoDE (Max-endorsement)	0.3269	0.1404	0.2938	0.7333	1.25	4.51

form with fixed (α, τ) as in Section 3.2, and only the contextual support score fed into Q_i changes.

Mean gating replaces p_i^+ with $\bar{p}_i^+ = \frac{1}{2} \sum_{j \in \{1,2\}} \mathcal{P}_{i,j}^+$ to aggregate endorsements by averaging rather than selecting the strongest signal. In learned gating, we replace p_i^+ with $\mathbf{w}_i^\top [\mathcal{P}_{i,1}^+, \mathcal{P}_{i,2}^+]^\top + b_i$, where $\mathbf{w}_i \in \mathbb{R}^2$ and $b_i \in \mathbb{R}$ are trained on BoLD.

Table 6 reports a consistent ordering across all metrics. Mean gating is the weakest variant because averaging contextual support dilutes a strong endorsement from one expert whenever the other provides no relevant cue. The lower gain of the learned gating indicates that $\mathcal{P}_{i,j}^+$ already encodes sufficient structure, and the additional linear layer causes overfitting rather than improving generalisation.

(6) Contribution of Facial Information: In this experiment, we ablate facial information to quantify its contribution to CA-MoDE. To do so, we filled the face region with black pixels across all database frames and retrained the architecture with parameters described in Section 4. Although face masking had little effect on person detection in $\mathcal{H}^{\text{object}}$, ERS decreased to 0.2635 on the test set. This substantial decrease demonstrates that facial expression serves as a complementary cue that reinforces body-based emotion recognition, even in a model designed primarily to exploit contextual and postural information.

6 CONCLUSION

Bodily expressions of emotion are inherently context-dependent. A raised fist in a stadium and a raised fist on a street may involve similar body configurations, but they invite different emotional interpretations. This ambiguity motivated us to propose an emotion recognition model that reasons not only about the body, but also about the contextual evidence surrounding it.

In this study, we introduced CA-MoDE, a context-aware mixture of domain experts for bodily emotion recognition in the wild. Unlike conventional approaches that incorporate scene and

object information as auxiliary feature augmentations, CA-MoDE represents context through emotion-context co-occurrence priors and applies them at the output-distribution level. Scene and object experts generate soft contextual pseudo-labels, while the proposed max-endorsement gate selects the strongest contextual support for each emotion dimension. This design reduces the dilution of informative contextual cues when one expert is uncertain or uninformative.

On BoLD, CA-MoDE achieved an ERS of 0.3269 using only single still images, outperforming the compared BoLD methods that rely on temporal cues. The improvement is mainly driven by stronger categorical emotion recognition, with gains in both mAP and mRA. At the same time, the lower VAD regression performance shows that continuous affective dimensions still require cues that static spatial context can provide only partially.

The ablation studies further support the central hypothesis of this work. Both scene and object context contribute to performance, although scene context provides the stronger signal. The max-endorsement gate outperforms mean and learned alternatives, suggesting that the co-occurrence priors already encode useful structure without requiring additional gating parameters. Finally, the face-masking experiment shows that facial information remains a complementary cue, even in a framework designed around body and context.

Overall, these findings indicate that spatial context is not merely background information for bodily emotion recognition. When modelled as structured probabilistic evidence, it can substantially improve recognition from still images and provide a complementary alternative to temporal modelling. Future work should investigate richer contextual priors, more reliable object evidence in low-resolution scenes, and hybrid models that jointly exploit spatial, facial, and temporal cues.

REFERENCES

- [1] H. Aviezer, Y. Trope, and A. Todorov. Body Cues, Not Facial Expressions, Discriminate Between Intense Positive and Negative Emotions. *Science*, 338(6111):1225–1229, 2012.

- [2] Amar D'Adamo, Marte Roel Lesur, Laia Turmo Vidal, Mohammad Mahdi Dehshibi, Daniel De La Prida, Joaquín R. Díaz-Durán, Luis Antonio Azpícueta-Ruiz, Aleksander Väljamäe, and Ana Tajadura-Jiménez. SoniWeight Shoes: Investigating Effects and Personalization of a Wearable Sound Device for Altering Body Perception and Behavior. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2024.
- [3] M. M. Dehshibi, Bitá Baiani, Gerard Pons, and David Masip. A Deep Multimodal Learning Approach to Perceive Basic Needs of Humans From Instagram Profile. *IEEE Trans. Affect.*, 14(2):944–956, 2023.
- [4] M. M. Dehshibi, T. A. Olugbade, F. Diaz-de Maria, N. Bianchi-Berthouze, and A. Tajadura-Jiménez. Pain Level and Pain-Related Behaviour Classification Using GRU-Based Sparsely-Connected RNNs. *IEEE Journal of Selected Topics in Signal Processing*, 17(3):677–688, 2023.
- [5] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.
- [6] Panagiotis Paraskevas Filntisis, Niki Efthymiou, Gerasimos Potamianos, and Petros Maragos. Emotion Understanding in Videos Through Body, Context, and Visual-Semantic Embedding Loss. In *European Conference on Computer Vision Workshops*, pages 747–755. Springer-Verlag, 2020.
- [7] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In *CVPR*, pages 6047–6056. IEEE, 2018.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.
- [9] Freek Hens, Mohammad Mahdi Dehshibi, Leila Bagheriye, Ana Tajadura-Jiménez, and Mahyar Shahsavari. LAST-PAIN: Learning Adaptive Spike Thresholds for Low Back Pain Biosignals Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 33:1038–1047, 2025.
- [10] A. Kendall, Y. Gal, and R. Cipolla. Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *CVPR*, pages 7482–7491. IEEE, 2018.
- [11] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed. Automatic Recognition of Non-Acted Affective Postures. *IEEE Trans. Syst. Man Cybern. Part B*, 41(4):1027–1038, 2011.
- [12] R. Kostí, J. Alvarez, A. Recasens, and A. Lapedriza. Context Based Emotion Recognition Using EMOTIC Dataset. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(11):2755–2766, 2020.
- [13] Xinpeng Li, Xiaojiang Peng, and Changxing Ding. Sequential Interactive Biased Network for Context-Aware Emotion Recognition. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–6. IEEE, 2021.
- [14] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, pages 740–755. Springer Cham, 2014.
- [15] Y. Luo, J. Ye, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang. ARBEE: Towards Automated Recognition of Bodily Expression of Emotion in the Wild. *International Journal of Computer Vision*, 128(1):1–25, 2020.
- [16] T. Mensink, E. Gavves, and C. G. M. Snoek. COSTA: Co-Occurrence Statistics for Zero-Shot Classification. In *CVPR*, pages 2441–2448. IEEE, 2014.
- [17] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *WACV*, pages 1–10. IEEE, 2016.
- [18] Jorge Ortigoso-Narro, Fernando Diaz-de Maria, Mohammad Mahdi Dehshibi, and Ana Tajadura-Jiménez. L-SFAN: Lightweight Spatially Focused Attention Network for Pain Behavior Detection. *IEEE Sensors Journal*, 25(10):18409–18418, 2025.
- [19] Ioannis Pikoulis, Panagiotis P. Filntisis, and Petros Maragos. Leveraging Semantic Scene Characteristics and Multi-Stream Convolutional Architectures in a Contextual Approach for Video-Based Visual Emotion Recognition in the Wild. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021.
- [20] G. Pons and D. Masip. Multitask, multilabel, and multidomain learning with convolutional networks for emotion recognition. *IEEE Trans. Cybern.*, 52(6):4764–4771, 2022.
- [21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9. IEEE, 2015.
- [22] Alexandros Xenos, Niki M. Foteinopoulou, Ioanna Ntinou, Ioannis Patras, and Georgios Tzimiropoulos. VLLMs Provide Better Context for Emotion Understanding Through Common Sense Reasoning. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2025.
- [23] H. X. Yu, W. S. Zheng, A. Wu, X. Guo, S. Gong, and J. H. Lai. Unsupervised Person Re-Identification by Soft Multilabel Learning. In *CVPR*, pages 2143–2152. IEEE, 2019.
- [24] Sitao Zhang, Yimu Pan, and James Z. Wang. Learning Emotion Representations From Verbal and Nonverbal Communication. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18993–19004. IEEE, 2023.
- [25] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2018.