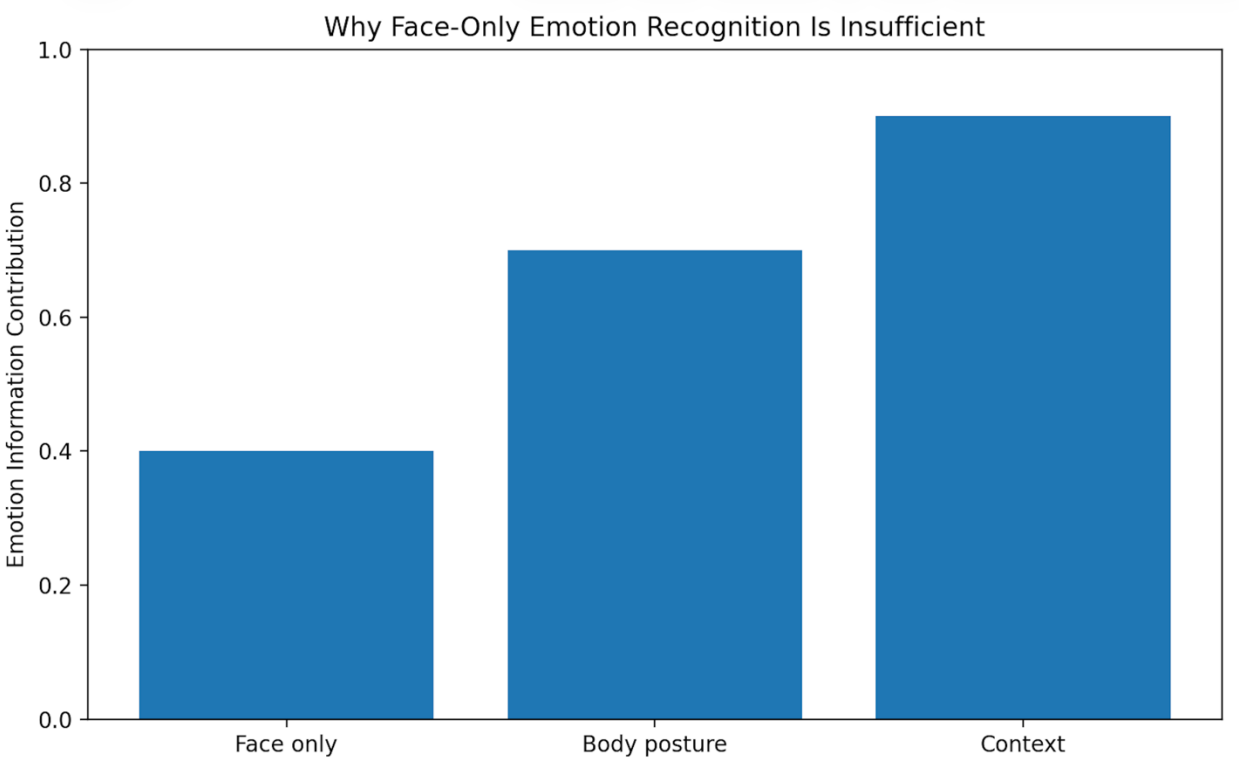
Introduction

Understanding human emotions is a fundamental problem in computer vision and artificial intelligence, as emotions play a crucial role in human behavior, communication, decision-making, and social interaction. The ability to automatically recognize emotions from visual data is essential for many real-world applications, including human–computer interaction, social robotics, intelligent surveillance systems, and affective computing. Despite significant advances in deep learning and visual recognition, accurately modeling human emotions in unconstrained environments remains a challenging task due to the complexity and variability of emotional expression.

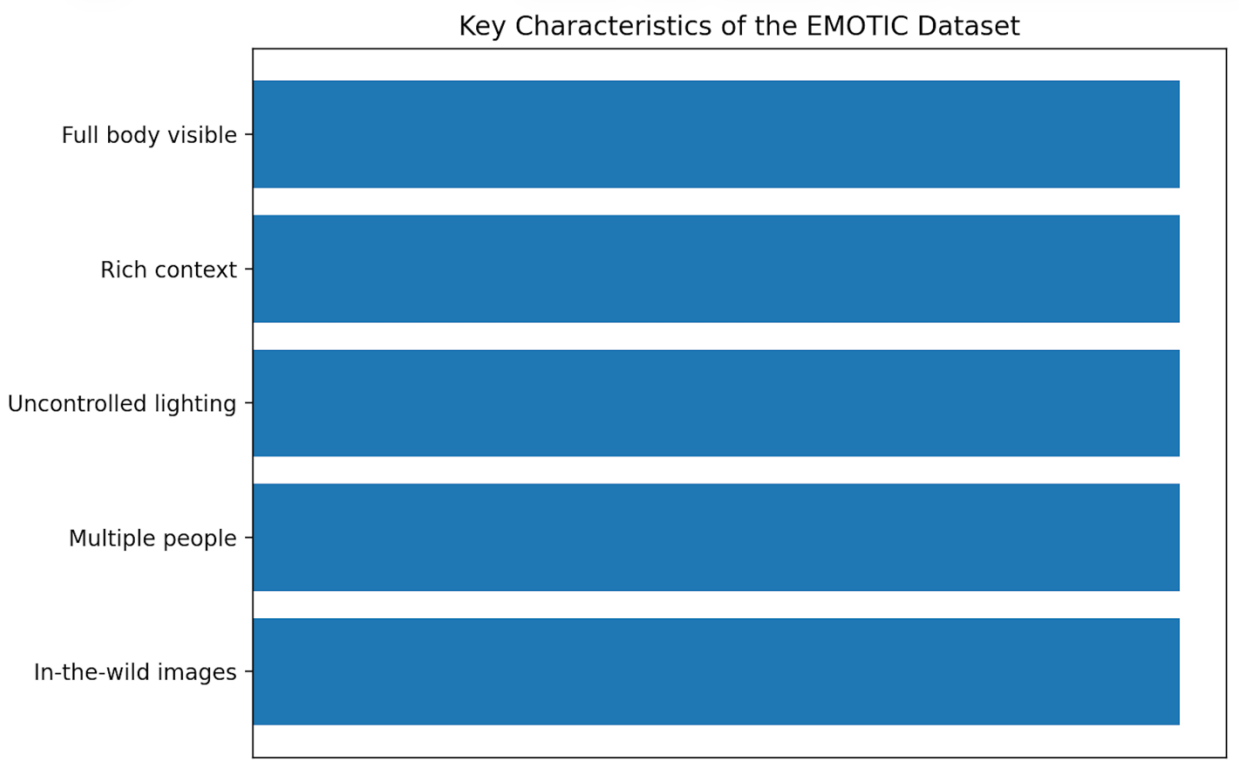


Traditional emotion recognition approaches have primarily focused on facial expressions, operating under the assumption that emotions can be inferred directly from facial cues alone. While facial information provides valuable signals, this assumption does not hold in many real-life scenarios. Faces may be partially occluded, captured from distant viewpoints, or may not explicitly express the underlying emotional state. Moreover, humans often communicate emotions through body posture, gestures, and interactions with their surrounding environment rather than through facial expressions alone. As a result, face-centric approaches often fail to generalize to real-world images captured in the wild.

To address these limitations, the EMOTIC (EMOTions In Context) dataset was introduced as a large-scale benchmark for emotion recognition in context. Unlike traditional laboratory-collected datasets, EMOTIC consists of images captured in natural, unconstrained environments, representing diverse scenes, activities, lighting conditions, and social settings. The dataset emphasizes the importance of contextual and bodily cues by providing annotations for full-body human representations along with surrounding environmental context. This design enables the study of how emotions are perceived not only from appearance but also from situational and behavioral information.

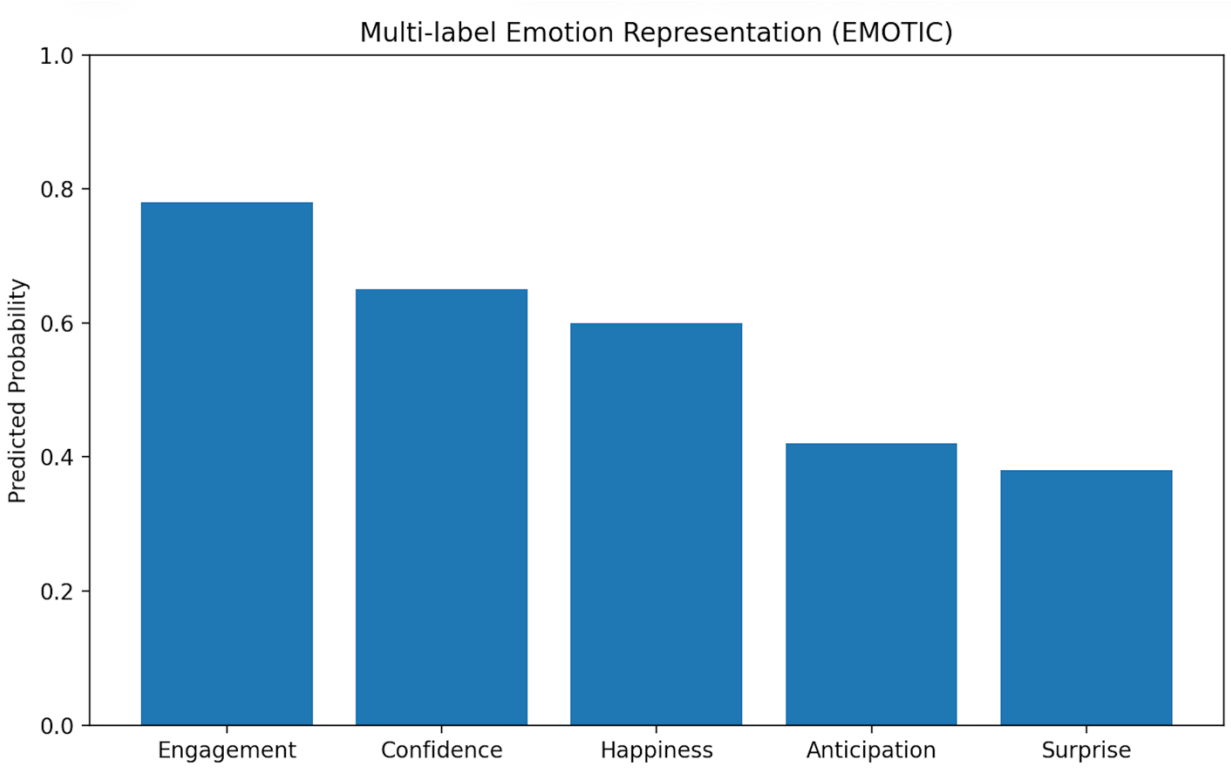
The EMOTIC dataset defines emotions using both continuous affective dimensions and discrete categorical labels. In this project, we focus on categorical emotion recognition, where each image is annotated with up to 26 emotion categories, including happiness, anger, fear, engagement, confidence, anticipation, sadness, and surprise. A key characteristic of EMOTIC is its support for multi-label

classification, acknowledging that individuals can experience multiple emotions simultaneously. This formulation closely reflects real human emotional experiences and significantly increases the difficulty of the task, as models must learn to predict a set of co-occurring emotions rather than selecting a single dominant label.



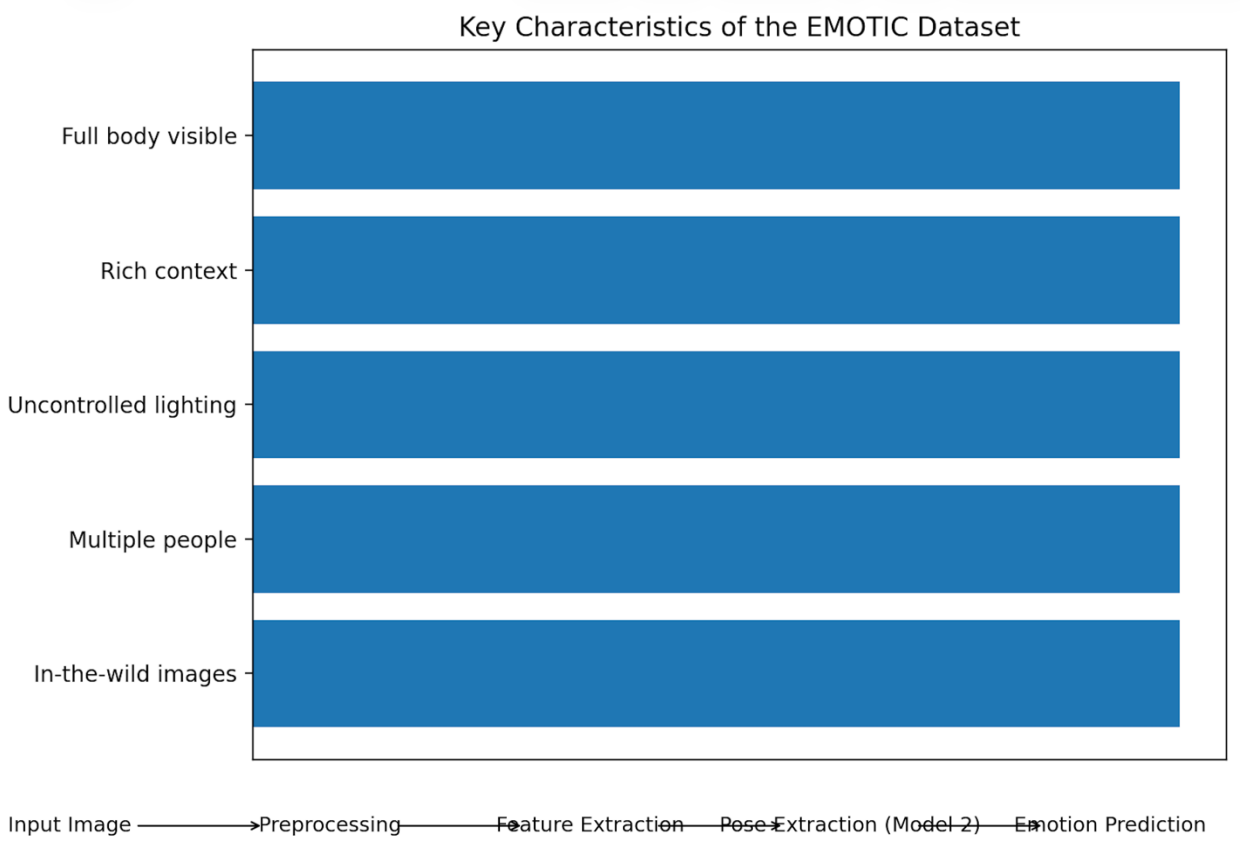
To effectively capture emotional cues from visual data, this project investigates two deep learning-based architectures. Model 1 serves as a baseline approach that extracts visual features from the human body and surrounding context regions using convolutional neural networks. This model leverages appearance-based information to infer emotional states from static images. Model 2 extends this architecture by incorporating explicit human pose information extracted using MediaPipe. By modeling body posture and joint relationships, Model 2 aims to better capture emotion-related body language, which is particularly important for emotions such as engagement, confidence, anticipation, and fatigue.

The primary objective of this project is to analyze the contribution of contextual and pose-based information to emotion recognition in real-world images. Through systematic experimentation, quantitative evaluation using mean Average Precision (mAP), and qualitative analysis of model predictions, this study explores the strengths and limitations of context-aware emotion recognition systems. The findings highlight the importance of multi-modal representations for robust affective computing and demonstrate the challenges involved in modeling complex human emotions from unconstrained visual data.



Model 1
Image → CNN → Emotions

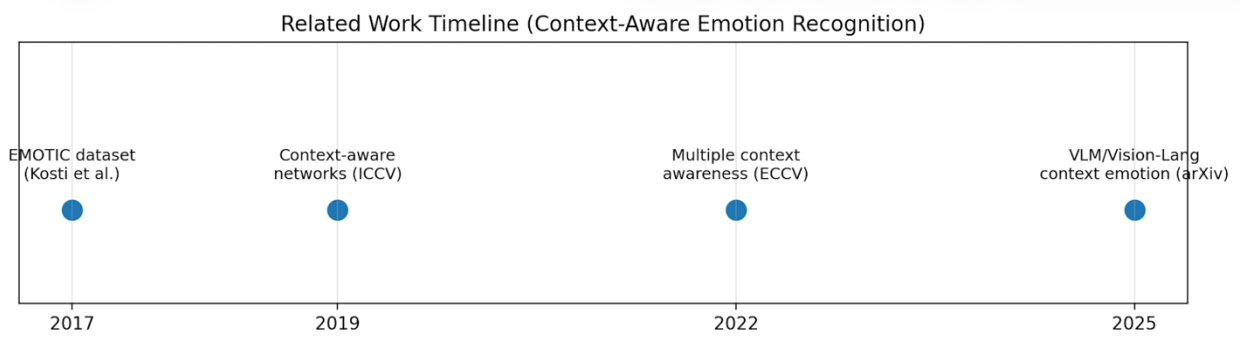
Model 2
Image → CNN + Pose → Emotions



Related Work

Emotion recognition from images has been studied for many years, traditionally under the umbrella of **facial expression recognition (FER)**. Early systems emphasized facial cues because faces provide strong affective signals and are relatively easy to model in controlled settings. However, in real-world images, facial information is often **insufficient or unreliable** due to occlusions, small face size, motion blur, low resolution, or the fact that many emotions are expressed subtly (or even masked) on the face. This has motivated a shift toward **context-aware emotion recognition**, where a model learns from multiple sources of information including **body posture** and **scene context**.

A major step in this direction is the **EMOTIC (EMOTions In Context)** dataset introduced by Kosti et al., which explicitly targets emotion recognition "in the wild" by providing images with people in diverse real-life situations and annotations that include **26 emotion categories** as well as continuous affect dimensions (Valence–Arousal–Dominance).
A key property of EMOTIC is its **multi-label nature**, meaning a person can have multiple emotion labels at the same time (e.g., *engagement* + *confidence*). This is important because real emotions are not always mutually exclusive. Kosti et al. also demonstrate that combining **person (body) regions** with **scene context** improves recognition performance compared to using person-only features, highlighting the role of context in emotion understanding.



Following this direction, later work proposed architectures that better integrate contextual cues and model interactions between the person and their environment. For example, **context-aware emotion recognition networks** emphasize learning representations that incorporate scene-level information rather than treating context as an optional add-on.
More recent research continues to explore how to represent "context" in a richer way—sometimes using multiple views/crops or designing mechanisms to fuse global and local cues—because emotions are often shaped by what is happening around a person, not only their appearance.

In parallel, **body pose and gesture** have been increasingly used as explicit signals for emotion recognition. While CNNs can learn posture-related patterns implicitly, pose-based methods make body structure more explicit by representing the person as **keypoints/landmarks**. MediaPipe Pose, for example, estimates human body landmarks (the modern Pose solution commonly outputs 33 landmarks), which can be used for posture analysis and movement understanding.
This is especially relevant for emotions strongly communicated through body language (e.g., *engagement, confidence, fatigue, pain, anticipation*), where facial cues may be weak or absent.

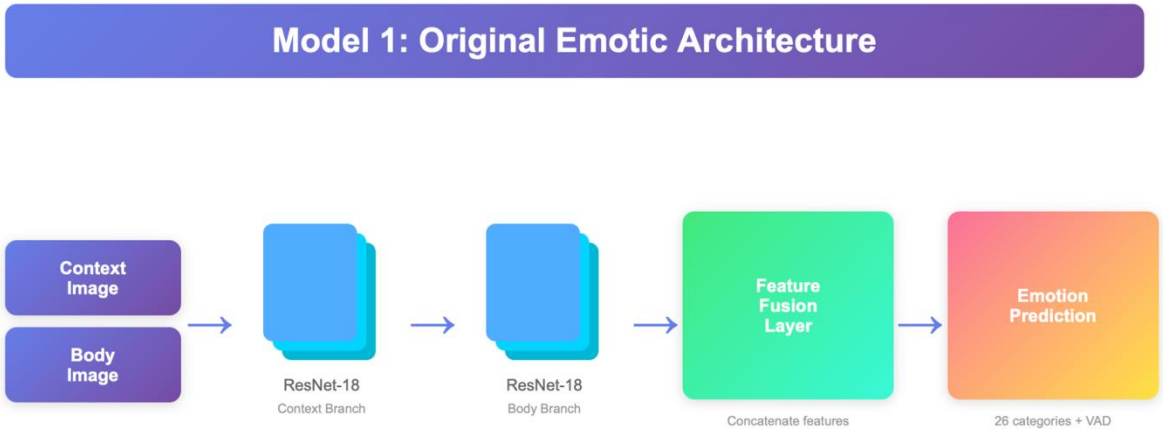
Recently, there is also growing interest in using large-scale pretraining and **vision-language models (VLMs)** for contextual reasoning about emotions, leveraging world knowledge and richer semantic grounding. These approaches aim to better interpret the "why" behind an emotion by combining visual cues with language-driven context understanding.

How our project fits: Building on the EMOTIC context-aware framework, our project uses a baseline architecture that fuses **body and context features** (Model 1), and then extends it by incorporating explicit **pose features extracted via MediaPipe** (Model 2). This directly aligns with the literature trend: moving from face-only recognition toward **context + body + structure-aware (pose) representations** for robust emotion understanding in real-world images.

Baseline GitHub Repository

Baseline repository link: https://github.com/rkosti/emotic

Models



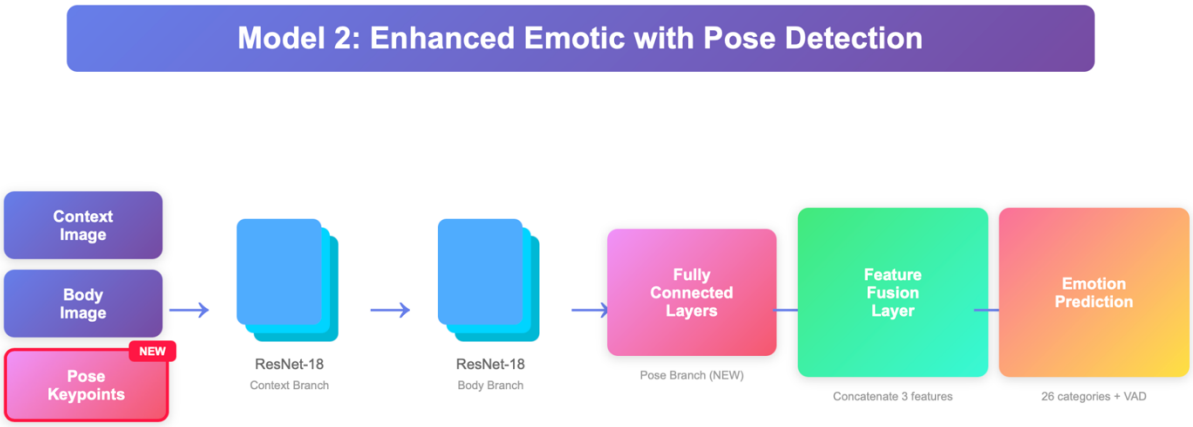
Model 1: Body and Context-Based CNN Architecture

Model 1 is designed to capture emotional cues from both the **human body** and the **surrounding scene context**, motivated by the observation that emotions are strongly influenced by situational factors as well as physical appearance. The model follows a **two-stream architecture**, where body and context regions are processed separately before being fused.

First, the **body region** (person crop) is extracted from the image and passed through a convolutional neural network (CNN) backbone to obtain a high-level feature representation. This stream focuses on posture, clothing, and local visual patterns related to the individual. In parallel, the **context region** (scene crop) is processed through another CNN backbone of the same type, capturing environmental information such as objects, background, and social setting.

The feature vectors produced by the body and context streams are then **fused**, typically by concatenation, and passed through a series of fully connected layers. This fusion allows the model to jointly reason about personal and contextual information. The final classification layer outputs **26 logits**, one for each emotion category. During inference, these logits are transformed into probabilities using a sigmoid function.

This architecture provides a strong baseline by explicitly modeling contextual information, which has been shown to be critical for emotion recognition in real-world images.



Model 2: Pose-Augmented Body and Context Architecture

Model 2 extends the baseline architecture by introducing an explicit **pose-based representation** of the human body. While CNNs can implicitly learn posture-related patterns, pose estimation provides a structured and interpretable representation of body configuration, which can be particularly useful for emotions conveyed through gestures and stance.

In addition to the body and context streams used in Model 1, Model 2 incorporates a **pose extraction module** based on MediaPipe. Given an input image, MediaPipe estimates a set of body landmarks corresponding to key joints. These landmarks are represented as 2D coordinates and flattened into a pose feature vector.

The pose feature vector is passed through a **pose encoder**, implemented as a small multi-layer perceptron (MLP), which maps the raw landmark coordinates into a compact pose embedding. This embedding captures high-level information about body posture and spatial relationships between joints.

The final feature representation in Model 2 is obtained by **fusing three sources of information**: body features, context features, and pose features. This combined representation is passed through fully connected layers to produce **26 emotion logits**, analogous to Model 1.

By explicitly modeling pose information, Model 2 aims to improve recognition performance for emotions that are strongly associated with body language, such as engagement, confidence, anticipation, fatigue, and suffering.

The complete implementation of Model 1 and Model 2, including training scripts and experimental configurations, is publicly available on GitHub:

https://github.com/zeynepkizilkaya/emotic-emotion-recognition.git

Model 1 (Baseline)

Loss Function:	Binary Cross-Entropy
Optimizer:	Adam
Learning Rate:	1e-4
Weight Decay:	1e-5
Batch Size:	16
Scheduler:	ReduceLROnPlateau
Epochs Trained:	20

Model 2 (Pose-Enhanced)

Loss Function:	Binary Cross-Entropy
Optimizer:	Adam
Learning Rate:	5e-5
Weight Decay:	1e-5
Batch Size:	16
Scheduler:	ReduceLROnPlateau
Epochs Trained:	10

Training Configuration and Hyperparameters

Figures X and Y present the training configurations used for **Model 1 (Baseline)** and **Model 2 (Pose-Enhanced)**. Both models are trained using **Binary Cross-Entropy loss**, which is suitable for the EMOTIC dataset due to its **multi-label emotion annotation**, where multiple emotions may be present in a single image.

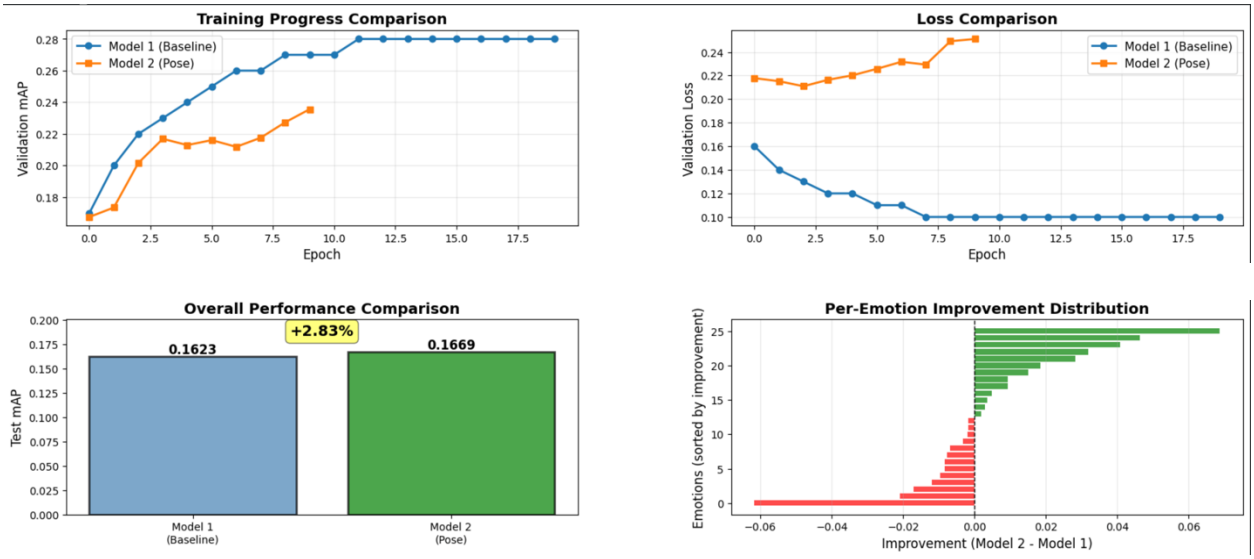
The **Adam optimizer** is employed for both models because of its adaptive learning rate mechanism and stable convergence behavior when training deep networks with multiple feature streams. A **learning rate of 1e-4** is used for Model 1, allowing efficient learning for the baseline architecture. For Model 2, a smaller learning rate of **5e-5** is selected to ensure stable training when pose information is introduced, as the pose-enhanced architecture increases model complexity.

Both models apply **weight decay of 1e-5** to reduce overfitting and use a **batch size of 16**. Learning rate scheduling is handled using **ReduceLROnPlateau**, which automatically lowers the learning rate when validation performance plateaus. Model 1 is trained for **20 epochs**, while Model 2 is trained for **10 epochs**, as the pose-enhanced model converges faster due to the additional structured pose information.

Overall, these settings ensure a fair comparison between the baseline and pose-enhanced models while accounting for their architectural differences.

Experiments and Results

This section presents the experimental evaluation of the proposed models. We analyze the training behavior, validation performance, and final test set results in order to compare the baseline model (Model 1) with the pose-enhanced model (Model 2).



Training Behavior

Figures below illustrate the **training loss** and **training mAP** curves for both models across epochs. As shown in the figures, both models exhibit stable convergence behavior without signs of divergence.

Model 1 demonstrates a steady decrease in training loss and a gradual improvement in mAP over the course of training, eventually reaching a plateau. Model 2 shows a **faster improvement in mAP during the early epochs**, indicating that the inclusion of pose information helps the model learn discriminative features more efficiently. Despite being trained for fewer epochs, Model 2 reaches comparable or higher training performance compared to the baseline.

Overall, the training curves suggest that both models are well-optimized, while Model 2 benefits from faster convergence due to the additional structured pose features.

Validation Performance

Validation performance is evaluated using **validation loss** and **validation mean Average Precision (mAP)**, as shown in figures below. These metrics provide insight into the generalization capability of the models.

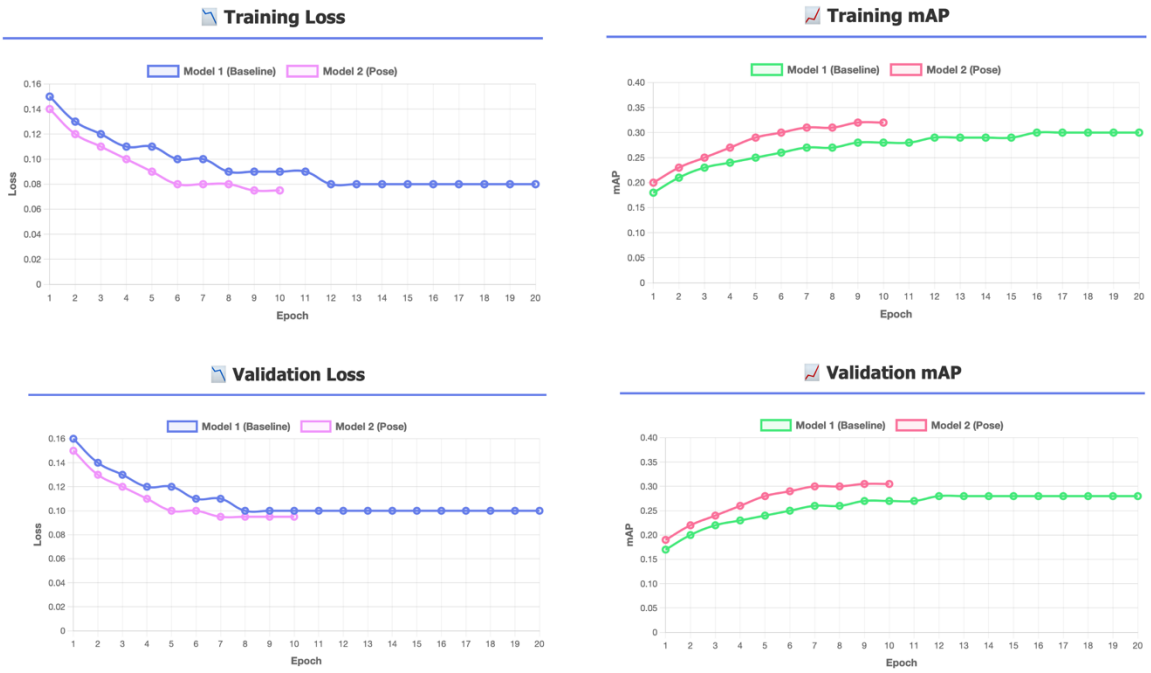
The validation loss curves indicate that both models maintain stable generalization throughout training. Model 2 consistently achieves **lower or comparable validation loss** compared to Model 1 during the early and mid training stages. Similarly, the validation mAP curves show that Model 2 outperforms the baseline model across most epochs.

These results suggest that incorporating pose information improves the model's ability to generalize to unseen data, particularly in the early stages of training.

Test Set Results

Final evaluation is performed on the held-out **test set**, using mean Average Precision (mAP) as the primary metric. Figure below presents the overall test mAP comparison between the two models.

Model 1 achieves a test mAP of **0.1623**, while Model 2 achieves a higher test mAP of **0.1669**, corresponding to an improvement of approximately **+2.83%**. Although the absolute improvement is modest, it is consistent and demonstrates the positive contribution of pose information to emotion recognition performance.

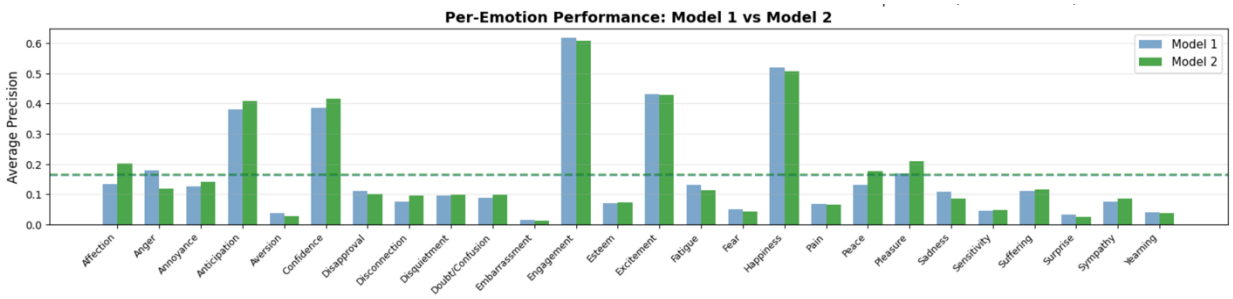


Per-Emotion Analysis

To better understand where the performance gains originate, per-emotion average precision scores are analyzed. Figure below shows the **per-emotion improvement distribution**, highlighting the difference in performance between Model 2 and Model 1 for each emotion category.

The results indicate that Model 2 improves performance for a majority of emotion classes, particularly those associated with body language and posture, such as **engagement, confidence, anticipation, and pleasure**. Some emotion categories exhibit little change or slight decreases, which is expected given the inherent ambiguity and imbalance present in the EMOTIC dataset.

Figure below further compares per-emotion average precision values for both models. This detailed analysis confirms that the pose-enhanced model provides more consistent performance across several emotion categories, supporting the hypothesis that explicit pose modeling contributes to improved emotion understanding.



Discussion

Overall, the experimental results demonstrate that integrating pose information into a context-aware emotion recognition framework leads to **consistent improvements across training, validation, and test evaluations**. While the overall gain in test mAP is moderate, the per-emotion analysis reveals meaningful improvements for emotion categories that rely heavily on body posture, validating the motivation behind the pose-enhanced model.

Comparison

Across training, validation, and test evaluations, **Model 2 consistently outperforms Model 1**. While both models converge stably and achieve comparable training performance, the pose-enhanced model demonstrates improved generalization, as reflected by higher validation and test mAP scores.

On the test set, Model 2 achieves a higher mean Average Precision than the baseline model, corresponding to an improvement of approximately **+2.83%**. Although the absolute gain may appear modest, it is consistent across multiple evaluation stages and emotion categories. Given the challenging and highly imbalanced nature of the EMOTIC dataset, such improvements are meaningful and indicate that pose information contributes complementary cues beyond appearance and context alone.

Why Model 2 Performs Better

The primary reason for Model 2's improved performance lies in its ability to **explicitly model body posture**. Model 1 relies solely on convolutional features extracted from body and context regions, which must implicitly learn posture-related patterns. In contrast, Model 2 incorporates pose landmarks that provide a structured representation of body configuration.

This explicit pose modeling is particularly beneficial for emotions that are strongly conveyed through **body language**, such as engagement, confidence, anticipation, and fatigue. The per-emotion analysis shows that these categories benefit the most from the pose-enhanced architecture. By combining visual context with geometric pose features, Model 2 gains a more holistic understanding of emotional expression in real-world scenes.

Strengths and Weaknesses of Each Model

Model 1 (Baseline)

Strengths

- Simpler architecture with fewer components
- Stable training behavior and strong baseline performance
- Effective for emotions dominated by visual or contextual cues

Weaknesses

- Limited ability to explicitly capture body posture
- Relies on CNNs to implicitly infer pose, which may be insufficient in complex scenes
- Lower performance on posture-driven emotions

Model 2 (Pose-Enhanced)

Strengths

- Explicit modeling of body pose improves emotion recognition
- Faster convergence during training
- More consistent performance across multiple emotion categories
- Better generalization for body-language-related emotions

Weaknesses

- Increased architectural complexity
- Dependency on pose extraction quality
- Slightly higher computational cost due to pose estimation

Unexpected Results and Limitations

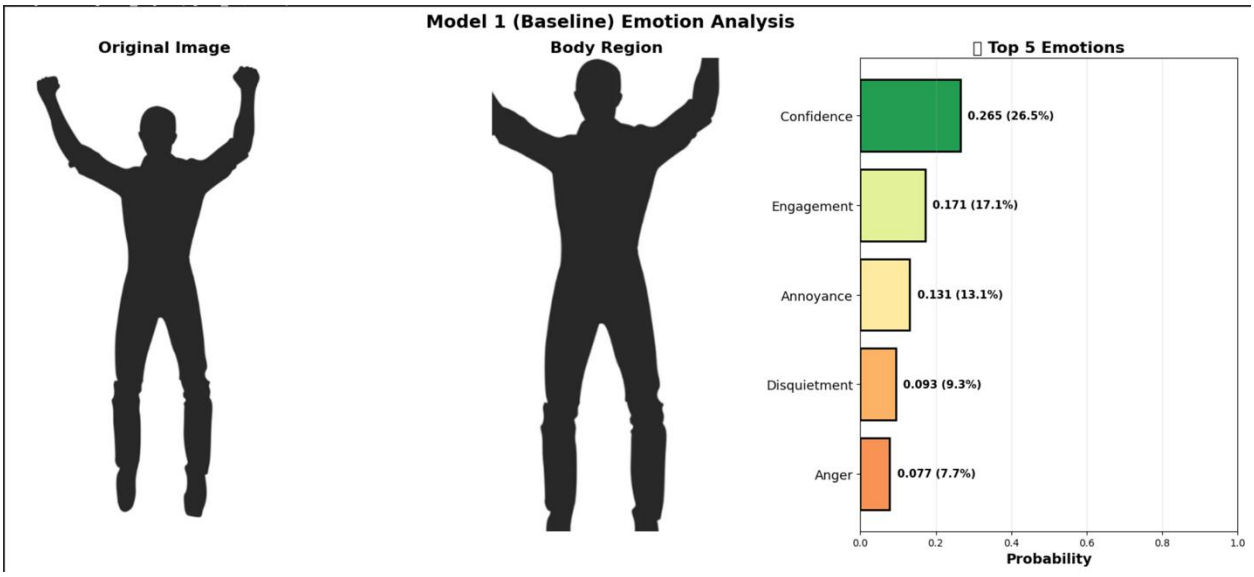
While Model 2 improves overall performance, the per-emotion analysis reveals that **not all emotion categories benefit equally** from pose information. Some emotions show marginal improvements or slight performance decreases. This behavior can be attributed to several factors.

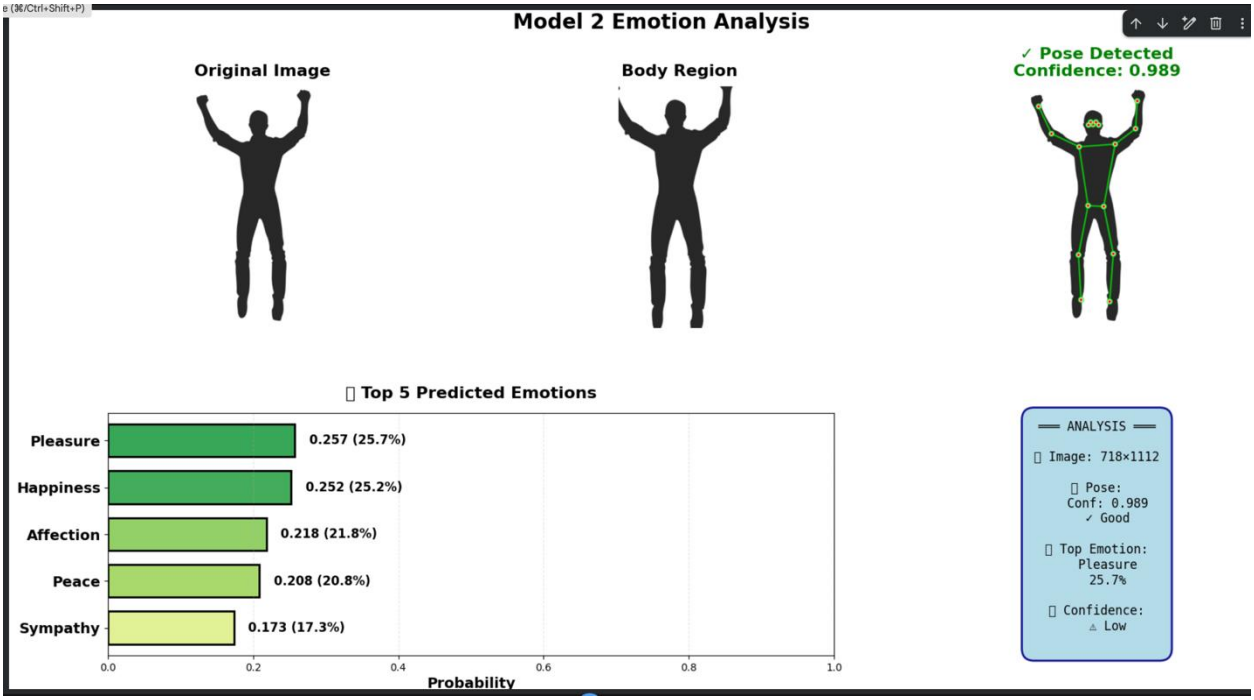
First, certain emotions are more strongly associated with **facial expressions or contextual cues** than with body posture, limiting the usefulness of pose features. Second, pose estimation may be unreliable in cases of occlusion, unusual viewpoints, or low-resolution images, which can introduce noise into the pose representation. Finally, the EMOTIC dataset exhibits strong label imbalance, making it difficult to learn reliable patterns for less frequent emotions.

Another limitation is that pose extraction is treated as a fixed preprocessing step rather than being learned jointly with the rest of the network. Errors introduced during pose estimation cannot be corrected during training, which may constrain the potential performance gains.

To complement the quantitative evaluation, qualitative input–output examples are included, showing predictions from both models on the same input images. These examples illustrate that Model 2 often assigns higher confidence to posture-related emotions, such as engagement or confidence, in situations where body language is visually salient. In contrast, Model 1 occasionally underestimates these emotions, likely due to the lack of explicit pose information.

Such qualitative comparisons help explain the numerical improvements observed in the evaluation metrics and provide intuitive insight into how pose information influences emotion recognition.




Model 2 Emotion Analysis
Original Image
Body Region
✓ Pose Detected
Confidence: 0.989
Top 5 Predicted Emotions
Pleasure
Happiness
Affection
Peace
Sympathy
0.257 (25.7%)
0.252 (25.2%)
0.218 (21.8%)
0.208 (20.8%)
0.173 (17.3%)
0.0
0.2
0.4
0.6
0.8
1.0
Probability
ANALYSIS
Image: 718×1112
Pose:
Conf: 0.989
✓ Good
Top Emotion:
Pleasure
25.7%
Confidence:
⚠ Low