

Cross-Level Scene-Agent Graph Learning for Context-Aware Multi-Label Emotion Recognition

K. Suresh^{1*}, L. Sherly Pushpa Annabel², T. Venketbabu³,
S. Hariharasudhan⁴, M.S.S. Sasikumar⁵, A. Ramar⁶

Department of Computer Science and Engineering, Chennai Institute of Technology, Chennai, India¹

Department of Artificial Intelligence and Data Science, St. Joseph's College of Engineering, Chennai, India²

Department of Computer Science and Engineering,

Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India^{3, 4, 5}

Department of Mathematics, Chennai Institute of Technology, Chennai, Tamil Nadu, India⁶

Abstract—In context-aware emotion recognition from images, it is necessary to consider the visual appearance of the target person along with the surrounding environment. In unconstrained settings, emotions may be simultaneously related to the target person, thus making the task multi-label. However, existing methods focus on either the representation of the target or the context without taking the interaction between the two into consideration. The current work proposes a scene-agent graph learning method for context-aware multi-label emotion recognition. An input image along with an annotation of the target person is passed through a pretrained ResNet-50 model to obtain 2048-dimensional visual representations of the scene and the target agent. Such representations are mapped to a 256-dimensional graph space where the scene and agent nodes are connected using an explicitly constructed relation structure and normalized graph propagation is performed. The proposed approach is validated on the EMOTIC dataset using the 26-category multi-label setting. The ablation study involves four different settings: target-only, scene-only, scene-agent reasoning, and full-fusion, using a consistent evaluation process. The Micro-F1 score of the target-only baseline is 0.260712, and its mAP is 0.256390, while the scene-agent reasoning approach improves Micro-F1 to 0.453180 and mAP to 0.275428. The best mAP is achieved by the full fusion setting with 0.278457, even though the Micro-F1 of 0.448038 is slightly lower than that of scene-agent reasoning. Category-level average precision, statistical paired analysis, error analysis at the sample level, and qualitative analysis provide additional insights into the performance of the proposed model.

Keywords—Context-aware emotion recognition; multi-label emotion recognition; scene-agent graph; graph learning; cross-level interaction; adaptive fusion; affective computing; EMOTIC

I. INTRODUCTION

Recognition of emotions from visual input is essential in the field of affective computing and human-centric systems. In unconstrained images, the emotional state of the target individual depends upon other individuals, actions, objects, and contextual elements around him/her. The EMOTIC dataset offers a naturalistic benchmark whereby the target individual receives multiple emotions.

The central challenge is one of incorporating target appearance information together with context information without assuming that they are separate sources of information.

Target-level features reflect information from the person himself/herself, while scene-level features are reflective of the surrounding visual environment. Modeling the relationship between them would be helpful.

Graph representation learning offers a natural way to perform relational reasoning since there is scope for propagation of information between related visual elements. For EMOTIC static image recognition, an actionable formulation involves clear representation of the scene and agent nodes, graph connectivity, feature dimensions, and inter-level interaction.

The current research in the domain of emotion recognition focuses on context and relation-based approaches. The key idea of the approach presented here is the incorporation of both scene-agent and agent-agent relations into a single graph with attention and fusion based on the target conditions for the recognition of emotions in static images based on the EMOTIC dataset with 26 categories of emotions.

The study employs a controlled four-configuration ablation consisting of A1 Target-Only, A2 Scene-Only, A3 Scene-Agent Reasoning, and A4 Full Fusion. The ablation setup facilitates isolation of scene knowledge, explicit relational reasoning, and fusion under the same EMOTIC protocol. Evaluation involves a combination of a threshold-dependent Micro-F1 measure and mAP and per-category AP measures that do not require any threshold, supported by paired statistical and qualitative analyses.

The major contributions include: 1) visual cross-level scene-agent graph for static image multi-label emotion classification; 2) reproducible 2048-to-256-dimensional space representation and relational interaction definition; 3) controlled A1-A4 ablation experiment design; 4) complementing Micro-F1, mean average precision (mAP), per-category average precision (AP), statistical, sample-wise, and qualitative analysis; and 5) analysis of metric behavior and study limitations.

The rest of the study is structured as follows. In Section II, related work on context-aware and graph-based emotion recognition is discussed. In Section III, the scene-agent graph learning approach is introduced, comprising visual feature extraction, graph building, graph reasoning, cross-level

*Corresponding author.

interaction, adaptive fusion, and multi-label prediction. In Section IV, the EMOTIC dataset, data processing procedure, experimental settings, model configuration, and implementation details are described. In Section V, evaluation metrics and statistical analysis are provided. In Section VI, the results of experiments are presented and analyzed. In Section VII, the limitations of the work are addressed, and Section VIII concludes the study.

II. RELATED WORK

A. Context-Aware Emotion Recognition

Context-based emotion recognition takes into account the visual appearance of the target and scene context. The EMOTIC dataset is used for multi-label emotion recognition in a realistic scenario [1]. Several works have studied environmental context, scene graphs, hypergraph relationships, and viewpoint-specific contexts [2]–[5]. There are group-level emotion recognition and visual relationship modeling to understand the interaction between humans and visual objects [6], [7]. Some other works deal with visual-text prompts and contrastive learning [8], [9]. Other works have addressed multimodal video emotion recognition, environmental context, cognitive assistance, and robust visual transformers [10]–[12]. Even though all of these works are helpful for understanding the background, their problems and inputs are different from static image-based EMOTIC recognition.

B. Scene, Agent, and Relational Modeling

Graph-based techniques provide a formulation for the interactions between individuals and visual entities. For example, Chen et al. have studied context-aware emotion recognition based on interactions between agents and scenes, utilizing ICENet, which is an approach that fuses agent-level facial and body posture characteristics with scene information [13]. Also, Gong et al. developed a hybrid fusion technique for emotion recognition at the group level under complex environments [14], and Steward et al. synthesized findings on face-context interactions in emotion perception via meta-analysis [15]. Such works encourage the consideration of interactions between target person characteristics and contextual visual data. The current framework includes scene nodes, target agent nodes, and other agent nodes in one graph, which is followed by propagation, target-agent attention, and adaptive fusion steps.

C. Multi-Label Emotion Recognition and Visual Representation Learning

Since EMOTIC involves many labels, as there could be many categories of emotions associated with one target person simultaneously, EMOTIC becomes a multi-label dataset. Visual emotion recognition literature has investigated various approaches to representation learning, such as visual transformer and attention-based selective fusion of emotions [16]; hierarchical co-consistency quantization and binary representation of emotions [17]; dynamic kernel approach for video-based facial expression recognition [18]; multi-task learning of both dimensional and categorical emotions [19]; facial-landmark-based and attention-based context modeling [20]; and multiscale dynamic global and static local attention

mechanisms [21]. In addition to that, some related research includes investigating robustness for in-distribution, out-of-distribution, and unconstrained environments [22]; healthcare-driven facial expression recognition [23]; and emotion recognition in human-robot interaction [24]. These works show different visual emotion recognition scenarios, but in our study, we concentrate on image-based, person-centered 26-category multi-label recognition of emotions based on contextual scene-agent data.

D. Multimodal Emotion Recognition and Fusion Strategies

Multimodal and fusion-based systems involve vision, sound, text, dialogue, or environment cues. Related papers apply methods including audio-visual attention, conversation transformer, situational attention, cross-modal enhancement, and cross-modal transfer learning [25]–[29]. Recent papers focus on memory-enhanced multimodal transformers, context-aware human-robot interaction, comprehensive multimodal survey, conversation agent assessment, contact-less multimodal recognition, and review on multimodal systems trends and challenges [30]–[32], [33], [34], [35]. Though the above methods involve attention and fusion of modalities, their modalities and task definitions are not the same as static image EMOTIC recognition. This study, however, employs visual scene and target-agent cues only.

E. Context, Causal Modeling, Robustness, and Explainability

Causal debiasing and de-confounded training methods have also been used for context-aware emotion recognition [36]. Explainability and robustness have been approached through topological data analysis and distribution-based evaluations [37]. This study makes use of relational models and does not use causal interventions or causal identification.

The dynamic and real-time environments involve reinforcement learning-based video emotion recognition, spatio-temporal models, fog-assisted real-time classification, and visual transformer robustness [38], [39], [40], [41]. The thorough review of the machine learning techniques used for recognizing human emotions is provided [42]. The above works serve a wider context; however, they do not cover the EMOTIC problem setting.

F. Recent Context-Aware Studies and Research Positioning

The recent context-driven approaches involve agent-scene interaction models, contrastive contextual emotion representations [43], and a review of context-driven emotion recognition [44]. They further affirm the ongoing attention to explicit representations of context and relations. It covers related streams of context-driven, relational, visual, multimodal, robust, and recent context-driven emotion recognition. This set of studies sets the wider research context of the current static image-based EMOTIC approach.

This study is framed under the domain of static images, person-oriented, multi-label emotion recognition on EMOTIC. It tests whether explicit scene-agent and agent-agent connections within a unified graph improve multi-label classification, without necessarily claiming numerical superiority over approaches based on other modalities or datasets.

III. PROPOSED DICG-AF FRAMEWORK

A. Overview

Our proposed Dual-Level Interactive Context Graph with Adaptive Fusion (DICG-AF) system uses context-aware multi-label emotion recognition from the EMOTIC dataset. The features for the scene and target agent are extracted using the pre-trained ResNet-50 and are projected to the graph space of dimension 256.

A unified graph is composed of one scene node, one target-agent node, and several other-agent nodes. Scene-agent and agent-agent interactions are modeled through normalized graph propagation for context exchange. Two graph convolution layers generate graph-based node representations. In A4, the target-agent representation is used for cross-level attention, followed by adaptive gated fusion and 26 independent sigmoid outputs.

The final representation of 256 dimensions is then converted into 26 probability outputs for emotion, following the EMOTIC multi-label formulation. Thus, the entire process takes place in the following order: Input Image and Target Annotation → ResNet-50 Feature Extraction → Projection from 2048 to 256 Features → unified Scene-Agent Graph Construction → 2-layer Graph Convolution → Extraction of Scene and Target-Agent Nodes → Agent-Scene Attention Mechanism → Adaptive Gated Fusion → 26-class Multi-label Classification.

The operational details are described in Sections III-B to III-K and depicted in Fig. 1 and 2. The architecture is depicted in Fig. 1. The framework being analyzed utilizes image-only features; audio, dialogue, text, video, temporal, and causal-intervention branches have not been used in the experiments reported.

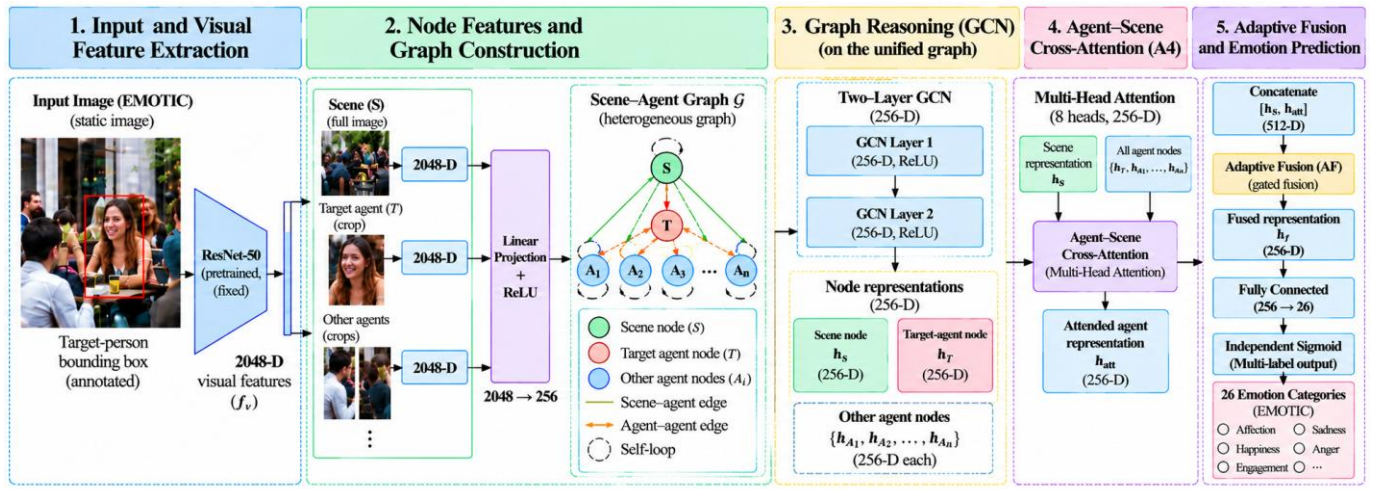


Fig. 1. Overall architecture of the proposed DICG-AF framework for context-aware multi-label emotion recognition.

B. Problem Formulation

Let an RGB image be represented as

$$X \in \mathbb{R}^{H \times W \times 3} \quad (1)$$

where, H and W represent the height and width of the image, respectively. Let b_t represent the annotated bounding box of the target person t . The goal is to predict the emotion classes of this particular target person t from the combination of target-level and contextual information from the image.

The target label in EMOTIC is encoded using a 26-dimensional multi-hot vector:

$$Y \in \{0, 1\}^{26} \quad (2)$$

For any specific target individual, there can be multiple elements of Y having the value one at the same time. It means that the task is to estimate probabilities for each emotion class independently:

$$pk = P(Yk = 1 | X, b), k = 1, \dots, 26 \quad (3)$$

The final probability vector is

$$p = [p_1, p_2, \dots, p_{26}] \in (0, 1)^{26} \quad (4)$$

Unlike a single-label classifier, which is mutually exclusive, this proposed classifier does not utilize a softmax output. Sigmoid functions are applied independently so that different emotions can be classified simultaneously.

C. Visual Feature Extraction

The pre-trained ResNet-50 model is used as the feature extractor. The classifier head is removed from the model to obtain an embedding vector with 2048 dimensions. During the graph learning tasks, the backbone remains fixed to ensure that the contribution of relational reasoning, cross-level interaction, and adaptive fusion can be assessed properly.

In each target-person example, there are two visual forms that are captured. The first one captures the entire context of the image, whereas the other one captures the target-person part that has been extracted using the bounding box annotation. Hence, both scene context information and target-agent information are captured. Let the feature extracted from the scene context be

$$f_s \in \mathbb{R}^{2048} \quad (5)$$

and the feature corresponding to an individual agent be

$$f_a \in \mathbb{R}^{2048} \quad (6)$$

For an image having N_s scene entities and N_a persons, the respective feature matrices are expressed as

$$X_s \in \mathbb{R}^{N_s \times 2048} \quad (7)$$

and

$$X_a \in \mathbb{R}^{N_a \times 2048} \quad (8)$$

The 2048-dimensional features are then mapped to the 256-dimensional graph representation space. As such, this mapping ensures that the cost of propagation in the graph is lowered while keeping the same dimension for the representations.

D. Scene-Agent Graph Construction

To combine global scene context and person-level interactions, the proposed framework constructs a unified scene-agent graph $G = (V, E)$ using the visual embeddings obtained for every EMOTIC image. In particular, the graph comprises one scene node S , one target-agent node T , and n_a additional agent nodes corresponding to other annotated persons present in the image. Therefore, the total number of nodes in the graph becomes $N = 2 + n_a$. The scene node represents global contextual information, whereas the target and other agent nodes represent person-level visual information.

Initialization of the node feature matrix is done by stacking the projected 256-dimensional visual features of the scene and people instances:

$$H^{(0)} = [h_s, h_t, h_{A_1}, h_{A_2}, \dots, h_{A_{n_a}}]^T \quad (9)$$

where, $H^{(0)} \in \mathbb{R}^{N \times 256}$. The scene feature can be derived from the entire image, whereas the target-agent and other-agent features are derived from the person regions. The unified graph facilitates the propagation of context-based and interpersonal information within a single relationship framework.

Two kinds of relational edges are introduced. Scene-agent edges relate the scene node S with person nodes, making global context information and person-specific information work together in the graph. Agent-agent edges relate person nodes to describe the relationships among persons in the same image. Self-loops are also considered to make sure that each node keeps its own representation in the process of graph propagation. The resulting graph can be denoted as $G = (V, E)$, where V represents the scene and agent nodes and E represents the scene-agent and agent-agent relationships.

The adjacency matrix is then symmetrically normalized by adding self-loops. Let A denote the adjacency matrix obtained after adding self-loops and I denote the identity matrix. The normalized adjacency matrix is expressed as

$$\hat{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} \quad (10)$$

where, $\tilde{A} = A + I$ and $\tilde{D}$ is the diagonal degree matrix of $\tilde{A}$. It ensures that information is propagated consistently through nodes with different numbers of neighboring nodes.

E. Agent Interaction Representation

$T, A_1, \dots, A_{n_a}$ The target node is separated from the other agent nodes. The scene-agent edges represent the global context, while the agent-agent edges represent the interpersonal context, and graph reasoning generates the target-agent embedding used in prediction and A4's attention mechanism.

Agent-to-agent links enable information about other seen agents to be used for target representation. The target agent representation is obtained from the graph reasoning on the target node, whereas the scene node retains the global context.

F. Graph Representation Learning

The graph reasoning takes place on the scene-agent graph through two consecutive graph convolution layers. The node representation for the l -th graph convolution layer is calculated as

$$H^{(l+1)} = \text{ReLU}(\hat{A}H^{(l)}W_l) \quad (11)$$

where, $H^{(l)}$ is the matrix representing the node features at the l -th layer, $\hat{A}$ represents the symmetrically normalized adjacency matrix, W_l represents the projection matrix, and ReLU represents the rectified linear activation function. The first graph-convolution layer operates on the 256-dimensional projected node features and produces updated 256-dimensional representations. The second graph-convolution layer propagates information over the relational structure, keeping the dimensions of the feature the same. Hence, the final graph representation is

$$H^{(2)} \in \mathbb{R}^{N \times 256} \quad (12)$$

Since the graph includes both scene and agent nodes, each node representation incorporates information from its respective contextual and interpersonal neighbors. The final representation of the scene is acquired from the scene node,

$$h_s = H_s^{(2)} \quad (13)$$

since the target-agent representation is extracted from the target node,

$$h_t = H_t^{(2)} \quad (14)$$

and the other node representations,

$$\{h_{A_1}, h_{A_2}, \dots, h_{A_{n_a}}\} \quad (15)$$

retain information from the other persons and are subsequently available to the agent-scene attention mechanism in the full-fusion configuration.

G. Scene and Target-Agent Representation Extraction

Once the graph propagation is done, the scene and target-agent nodes are explicitly extracted from the shared graph representation.

$$h_s = H_s^{(2)} \in \mathbb{R}^{256} \quad (16)$$

and the target-agent representation is defined as,

$$h_t = H_t^{(2)} \in \mathbb{R}^{256} \quad (17)$$

All these representations are already in the same dimensionality since all the nodes in the graphs are mapped to the 256-dimensional graph space before graph reasoning takes place. Hence, there will be no need for any additional dimensionality reduction for graph-level representation at this stage. The other representations related to agent-nodes are kept for the next cross-level attention. This node-level selection also ensures that there is a direct mapping between the representations that are used in the final prediction with respect to their positions in the unified scene-agent graph. The resulting h_s and h_T representations are passed to the cross-level scene-agent interaction module described in Section III-H.

H. Cross-Level Scene-Agent Interaction

In A4, the target-agent representation queries the entire graph using eight-head attention; then comes adaptive gated fusion to fuse the attended target and scene representations.

The application of multi-headed attention is done by an eight-headed mechanism in the graph representation space of 256 dimensions. The attention operation is thus represented.

$$h_{att} = \text{MHA}(h_T, H^{(2)}, H^{(2)}) \quad (18)$$

where, $h_{att} \in \mathbb{R}^{256}$ is the attended target-agent representation. Since the unified graph representation is used as the key-value context, the attention mechanism is capable of using information from the scene node in addition to the target and neighboring agent nodes. This attended representation is added to the target-agent representation to get the interacted target representation:

$$h_{int} = [h_T; h_{att}] \quad (19)$$

The target-agent and attended representations are concatenated together to create a vector of dimension 512. The learned linear projection transforms this vector into another vector with 256 dimensions before fusion.

This is followed by merging the target-agent interaction representation with the scene node representation using the adaptive gated fusion approach presented in Section III-I. The fused representation is finally fed to the multi-label classification head with independent sigmoid output of 26 dimensions. Fig. 2 provides an illustration of graph construction, graph reasoning, target-agent attention, and adaptive fusion.

It can be seen from Fig. 2(a)–(c) that the network keeps a consistent scene-agent relational structure, where the scene node, target-agent node, and other agent nodes are handled together. Scene-agent edges link the global context information to individual persons, while the agent-agent edges denote the relationship between persons. These two graph convolution layers perform message passing within this unified structure, and afterwards, the scene-node and target-agent features are explicitly extracted. It is also seen in Fig. 2(d) that the target-agent feature serves as the query in the eight-head multi-head attention layer, and the full graph feature acts as the key-value information. This unified graph is kept during the relational processing, and attention/fusion works on the obtained 256-dimensional features.

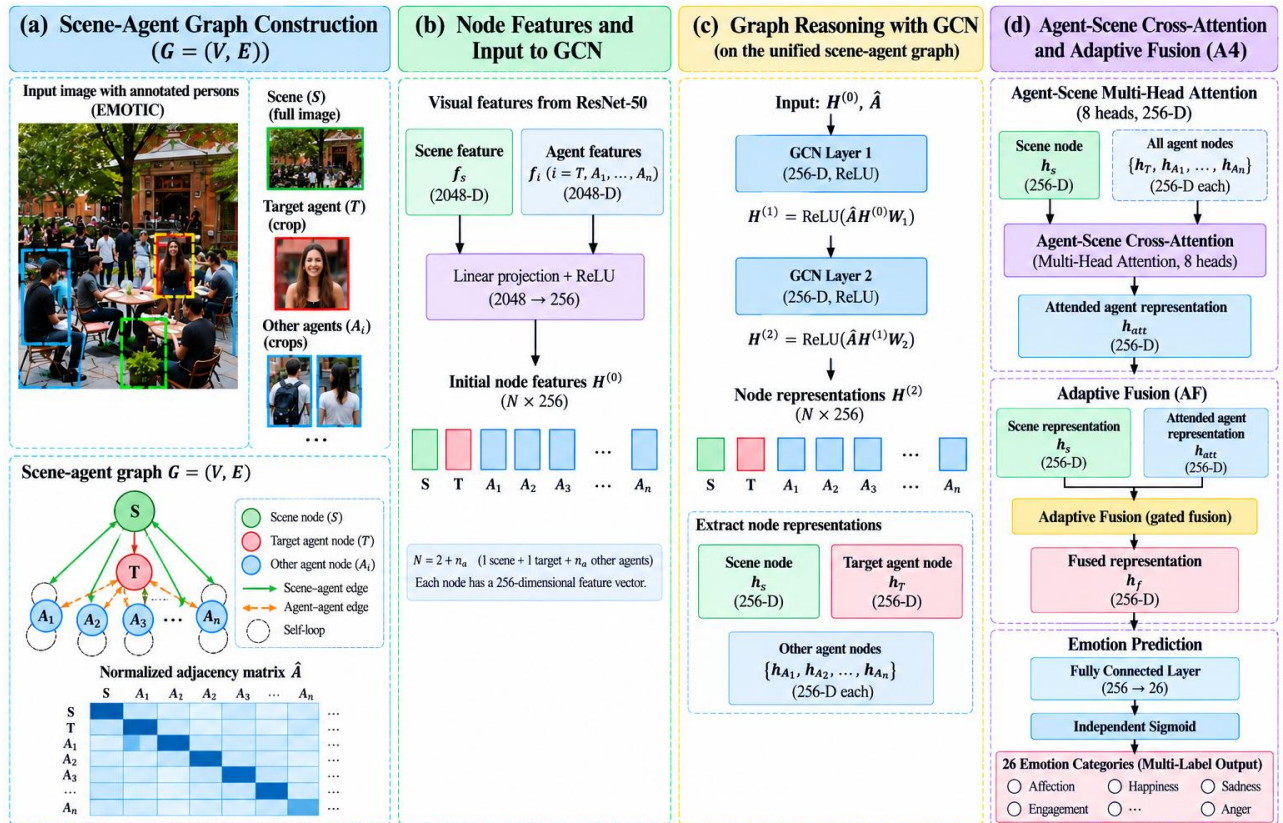


Fig. 2. Unified scene-agent graph construction, graph reasoning, target-agent attention, and adaptive fusion in the proposed framework.

I. Adaptive Fusion

After cross-level scene-agent interaction, the representation of the target-agent that is attended is fused with the scene representation using an adaptive gated fusion process. If $h_s \in \mathbb{R}^{256}$ is the representation of the scene-node and $h_{int} \in \mathbb{R}^{256}$ is the representation of the target-agent after interaction with the graph-level context, then

$$h_c = [h_s; h_{int}] \in \mathbb{R}^{512} \quad (20)$$

A learnable fusion gate is used to control the influence of the combined representation after transformation on the resulting fused representation. This gate comprises a nonlinear mapping consisting of two layers and a sigmoid activation function:

$$g = \sigma(W_{g2} \text{ReLU}(W_{g1}h_c + b_{g1}) + b_{g2}) \quad (21)$$

where, W_{g1} projects the 512-dimensional input space to a 256-dimensional space and W_{g2} preserves the 256-dimensional representation. Consequently,

$$g \in (0, 1)^{256} \quad (22)$$

In parallel, the concatenated representation is transformed into the 256-dimensional fusion space through

$$h_p = \text{LN}(W_p h_c + b_p) \quad (23)$$

where, $W_p \in \mathbb{R}^{256 \times 512}$ and LN denotes layer normalization.

The final fused representation is obtained using gated residual fusion:

$$h_f = g \odot h_p + (1 - g) \odot h_{int} \quad (24)$$

where, $\odot$ denotes element-wise multiplication. This way, the adaptive fusion process enables the model to keep important target-agent information and integrate scene-based information using the learned gate. This leads to the formation of $h_f \in \mathbb{R}^{256}$, this representation, which is further forwarded to the multi-label emotion classification head.

J. Multi-Label Emotion Classification

The fused output representation h_f is passed to a fully connected classification layer which generates a logit score for each of the 26 emotion categories described in EMOTIC:

$$o = W_c h_f + b_c \quad (25)$$

where, $W_c \in \mathbb{R}^{26 \times 256}$, $b_c \in \mathbb{R}^{26}$, and $o \in \mathbb{R}^{26}$. Because in EMOTIC, more than one emotion class can be assigned to a target individual at once, the 26 outputs are regarded as independent binary prediction tasks instead of being mutually exclusive categories. The probability for category k is computed independently by a sigmoid function:

$$p_k = \sigma(o_k) = \frac{1}{1 + \exp(-o_k)}, \quad k = 1, \dots, 26 \quad (26)$$

The complete prediction vector is

$$p = [p_1, p_2, \dots, p_{26}] \in (0, 1)^{26} \quad (27)$$

A fixed threshold value of 0.5 is applied for decision making in each of the categories:

$$\hat{y}_k = \begin{cases} 1, & p_k \geq 0.5, \\ 0, & p_k < 0.5. \end{cases} \quad (28)$$

The above independent-threshold approach enables predicting more than one category of emotion for a single target individual and hence conforms to the multi-label approach adopted in EMOTIC. The final evaluation process involves setting the threshold at 0.5 to transform the sigmoid probability to a binary prediction.

K. Multi-Label Training Objective

The training objective employs weighted binary cross-entropy with logits. Denote by $y_k \in \{0, 1\}$ the true label for emotion category k , and denote by o_k the associated model logit. Then, the weighted binary cross-entropy objective is given by

$$\mathcal{L} = -\frac{1}{26} \sum_{k=1}^{26} [w_k^+ y_k \log \sigma(o_k) + (1 - y_k) \log(1 - \sigma(o_k))] \quad (29)$$

where, w_k^+ is the positive-class weight for category k .

The weights for positive classes are computed using the following equation:

$$w_k^+ = \frac{N_k^-}{\max(N_k^+, 1)} \quad (30)$$

where, N_k^+ and N_k^- represent the numbers of positive and negative training examples in class k , respectively. In order to avoid extremely high weight values, the computed value is limited to the interval of $[1, 20]$.

Implementation is carried out through the use of BCEWithLogitsLoss with category-wise positive weighting, while logits are used for optimization and sigmoid probabilities for validation.

L. Computational Complexity

Where N , E , and d represent the number of graph nodes, graph edges, and representation dimensions, respectively, with $d = 256$. The initial projection from 2048 to 256 dimensions has a complexity of $O(N \cdot 2048 \cdot 256)$. A single sparse graph-convolution layer will have $O(Ed)$ cost for neighbor aggregation and $O(Nd^2)$ cost for the learning-based transformation; Therefore, the two-layer graph-reasoning module would be at $O(Ed + Nd^2)$ complexity level. With d , constant, this cost would scale linearly with the number of graph edges.

The eight-head attention blocks work with the 256-dimensional features, with their computational complexity defined by the attention model used. The interaction cost of full self-attention is $O(N^2d)$, while sparse attention complexity is influenced by the graph edges involved. The complexities of adaptive fusion and classification blocks are $O(N \cdot 512 \cdot 256)$ and $O(N \cdot 256 \cdot 26)$, respectively. In general, the proposed framework utilizes feature projection, sparse graph reasoning, attention fusion, and classification, where the complexity of graph message passing can be regarded as almost linear, while the full self-attention introduces a quadratic interaction cost term.

IV. EXPERIMENTAL SETUP

A. Dataset

The framework is evaluated on the EMOTIC (EMOTions in Context) dataset, which serves as a naturalistic benchmark for context-based emotion recognition [1]. The 26 categories are considered independent labels.

The 26 emotion labels are as follows: Affection, Anger, Annoyance, Anticipation, Aversion, Confidence, Disapproval, Disconnection, Disquietment, Doubt/Confusion, Embarrassment, Engagement, Esteem, Excitement, Fatigue, Fear, Happiness, Pain, Peace, Pleasure, Sadness, Sensitivity, Suffering, Surprise, Sympathy, and Yearning. Since one subject can have more than one emotion at a time, the problem is defined as a multi-label problem with 26 dimensions instead of a mutually exclusive single-label problem. The original organization contains 23,554 images (17,077 train, 2,088 validation, and 4,389 test). The present study uses person-level annotations and the final valid sample counts reported in Table I.

TABLE I. FINAL EMOTIC DATASET STATISTICS

Split	Person-Level Samples	Mean Labels per Sample
Training	23,705	1.9559
Validation	3,334	6.1569
Test	7,279	4.4161
Total	34,318	—

B. Data Preparation and Quality Control

There was a bounding-box and annotation audit prior to feature extraction. Corrections were made to invalid coordinate boxes where a valid area was possible, while one annotation with invalid coordinates and one without a class were removed, resulting in a total of 34,318 person samples.

Another corrupted target-person annotation was omitted due to the fact that the bounding box of this annotation extended outside the image boundaries. An annotation of an individual without any category was omitted since the official parser demands a category annotation.

C. Target Representation

Each test sample comprises an image, target person's bounding box, and 26-dimensional multi-hot emotion vector. The ground truth representation for a target person i is defined as

$$y_i = [y_{i1}, y_{i2}, \dots, y_{i26}]^T \quad (31)$$

where, $y_{ik} \in \{0, 1\}$. If the value is one, then the relevant emotion category is annotated for the target individual; otherwise, if it is zero, then the relevant emotion category is not annotated. This format allows for multiple emotion categories to be simultaneously annotated. As such, the experiment is not concerned with making the different emotion categories mutually exclusive; thus, no softmax output is used.

D. Visual Feature Extraction

The pretrained ResNet-50 model was kept fixed for all settings. The classification layer in this model was stripped off to generate 2048-dimensional features for the scene and target agent using the entire image and the target person region.

$$f_s \in \mathbb{R}^{2048} \quad (32)$$

for the scene and

$$f_a \in \mathbb{R}^{2048} \quad (33)$$

for the target agent.

These features were projected into a 256-dimensional space and fed into the standard graph-based pipeline. The backbone architecture was kept constant so that the effects of ablation are based on the considered scene, agent, graph, and fusion architectures.

E. Graph Construction and Model Configuration

In each graph sample, there is one scene node, one target agent node, and several other agent nodes. Scene-agent connections and agent-agent connections are concatenated together with self-loops and symmetric normalization prior to two 256-dimensional graph convolutional layers.

Edges between scene-agent link global context to person nodes; edges between agents link persons. Self-loops retain node-specific information when normalized propagation is performed. After two iterations of graph convolution, the nodes corresponding to scenes and target agents are extracted. While A3 uses these nodes directly without attention/fusion, A4 includes target agent attention with adaptive gating and fusion.

For the A3 Scene-Agent Reasoning architecture, the scene and target agent representations based on the graph reasoning are applied for prediction without explicit cross-level attention and adaptive fusion stages. In case of the A4 Full Fusion architecture, the target agent representation serves as a query in an eight-head multi-head attention with the full graph representation serving as a key-value. The resulting attended target representation is then merged with the scene representation by means of the adaptive gated fusion stage introduced in Section III-I. The resulting 256-dimensional representation is fed into a fully connected classification layer with 26 separate outputs. A sigmoid function is independently applied to each output to obtain multi-label emotion probabilities.

F. Training Protocol

All configurations use the identical data split, feature encoding, graph dimension, prediction dimension, and evaluation method. The graph is trained via the AdamW optimizer with a batch size of 128, a maximum of 30 epochs, and a weight decay of 10^{-4} . The A3 setup reported in the literature employs an initial learning rate of 5×10^{-4} , along with a ReduceLROnPlateau scheduler (factor 0.5, patience 2) and early stopping (patience 5). The model is optimized using weighted BCEWithLogitsLoss per category. The evaluation of performance is done at a fixed threshold of 0.5 for the separated test set.

G. Experimental Configurations

Table II shows the four controlled scenarios: A1 Target Only, A2 Scene Only, A3 Scene-Reasoning about the Agent, and A4 Complete Fusion. In the typical experimental procedure, scene information, relational reasoning, and cross-level fusion are isolated.

A1: Target-Only.

Target-person representation only.

A2: Scene-Only.

Scene representation only.

A3: Scene-Agent Reasoning.

Joint scene-agent graph reasoning without explicit cross-level fusion.

A4: Full Fusion.

Complete framework with scene representation, target-agent representation, graph reasoning, cross-level interaction, and adaptive fusion.

All the above four setups employ the final EMOTIC test set comprising 7,279 target person instances.

TABLE II. EXPERIMENTAL CONFIGURATIONS

Configuration	Scene	Target - Agent	Graph Reasoning	Cross-Level Interaction	Adaptive Fusion
A1: Target-Only	No	Yes	No	No	No
A2: Scene-Only	Yes	No	No	No	No
A3: Scene-Agent Reasoning	Yes	Yes	Yes	No	No
A4: Full Fusion	Yes	Yes	Yes	Yes	Yes

H. Reproducibility and Implementation Consistency

The process consists of a constant visual pipeline that involves: loading of EMOTIC data, dealing with target annotations, feature extraction with ResNet-50, projection to 256-dim from 2048-dim, graph building/ reasoning, cross-level interaction, adaptive fusion, and classification using a 26-category sigmoid loss function.

The pipeline was verified with real EMOTIC images before the main experiment. The final representations for testing consist of 7,279 scene features, 7,279 target-agent features, and corresponding target vectors 26 dimensions wide.

V. EVALUATION METRICS

A. Evaluation Protocol

Each sample provides 26 distinct probabilities. The threshold value used is a constant 0.5 for multi-label classification. Softmax accuracy is not employed since the classes are not exclusive to each other.

B. Precision and Recall

For each category, precision and recall are calculated using the true positives, false positives, and false negatives.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (34)$$

and

$$\text{Recall} = \frac{TP}{TP+FN} \quad (35)$$

where, TP , FP , and FN denote true positives, false positives, and false negatives, respectively. These measures are particularly relevant to EMOTIC because the 26 categories have different frequencies and multiple categories may be simultaneously active for a single target person.

C. Micro-F1 Score

Micro-F1 aggregates true positives, false positives, and false negatives across all 26 categories before computing F1.

$$F1_{\text{micro}} = \frac{2P_{\text{micro}}R_{\text{micro}}}{P_{\text{micro}}+R_{\text{micro}}} \quad (36)$$

where, P_{micro} and R_{micro} are the micro-averaged precision and recall. Micro-F1 gives a good description of the threshold-based multi-label prediction results while taking into account the overall balance of the positive and negative category classifications.

D. Average Precision

The Average Precision (AP) is calculated separately for each class based on the ranking of the probabilities of predictions. It is threshold-independent. For each class, AP characterizes the precision-recall tradeoff at varying threshold values.

E. Mean Average Precision

mAP is the arithmetic mean of the 26 category-wise AP values.

$$mAP = \frac{1}{26} \sum_{k=1}^{26} AP_k \quad (37)$$

mAP is especially well-suited for the EMOTIC approach since this metric takes into account all emotions and, at the same time, takes into account the multi-label characteristic of the task.

Micro-F1 and mAP are listed jointly because the former describes thresholded classification and the latter ranks items.

F. Per-Category Analysis

The AP per category is shown for all 26 categories to check if performance changes occur across the entire label set.

G. Statistical and Comparative Evaluation

Statistical comparisons were performed on the standard 7,279-sample test partition using paired samples. For each comparison, 10,000 pairs of bootstrap samples were created with two-sided 95% percentile confidence intervals and bootstrapped p-values.

The bootstrapped statistics presented in Table IV show the variation between the test samples, but do not measure the variation from training iterations or seed values. It is not known whether the difference in A3-A4 mAP is statistically significant (0.0030). Image-level dependence can also be a factor that influences uncertainty estimates because there could be more than one person in an image.

H. Overall Evaluation Criteria

Evaluation involves threshold-sensitive Micro-F1 score, threshold-insensitive per-category AP/mAP scores, and behavior results from ablation, statistics, and qualitative analysis.

VI. RESULTS AND DISCUSSION

A. Overall Ablation Results

Four different models were examined using the same EMOTIC test set comprising 7,279 samples. Table III shows the Micro-F1 and mAP results.

TABLE III. OVERALL ABLATION RESULTS ON THE EMOTIC TEST SET

Configuration	Micro-F1	mAP
A1: Target-Only	0.260712	0.256390
A2: Scene-Only	0.262044	0.261658
A3: Scene-Agent Reasoning	0.453180	0.275428
A4: Full Fusion	0.448038	0.278457

A1 provides the target-only baseline (Micro-F1 0.260712; mAP 0.256390), while A2 is similar in Micro-F1 (0.262044) and slightly higher in mAP (0.261658).

A2 shows that only scene information provides some thresholded performance, while ranking information is gained in comparison with A1.

A3 reaches Micro-F1 0.453180 and mAP 0.275428, corresponding to absolute gains of 0.192468 and 0.019038 over A1.

A4 reaches mAP 0.278457, 0.022067 above A1, while Micro-F1 is 0.448038, 0.005142 below A3. Thus, fusion changes the two aggregate metrics differently.

B. Contribution of Scene Context

Both A1 and A2 have small variations under the standard procedure. Scene information adds additional signals but is not enough alone for determining the emotions of the target person.

C. Effect of Scene-Agent Relational Reasoning

A3 increases Micro-F1 by 0.192468 and mAP by 0.019038 over A1, consistent with the contribution of explicit scene-agent relational reasoning.

The shift is mainly observed in the multi-label decision made at the selected operating point, whereas the graph propagation gives the joint scene-agent representation.

D. Effect of Adaptive Fusion

The mAP of A4 is improved by 0.003029 compared to A3, but the Micro-F1 drops by 0.005142, demonstrating metric-specific behavior of the adaptive fusion stage.

Paired bootstrap analysis supports these sample-level differences: A3-A1 and A4-A1 have $p < 0.001$, A2-A1 has $p = 0.313$, and A4-A3 has $p = 0.028$. Confidence intervals and effect sizes are reported in Table IV.

TABLE IV. PAIRED BOOTSTRAP ANALYSIS OF SAMPLE-LEVEL F1 DIFFERENCES

Comparison	Mean $\Delta F1$	95% Confidence Interval	Paired Effect Size	Bootstrap p-value
A2 - A1	+0.002003	[-0.001929, 0.006292]	0.012056	0.313
A3 - A1	+0.211472	[0.204773, 0.218142]	0.728006	<0.001
A4 - A1	+0.209452	[0.202947, 0.216050]	0.720809	<0.001
A4 - A3	-0.002020	[-0.003848, -0.000205]	-0.025175	0.028

E. Interpretation of Metric-Specific and Category-Level Behavior

The results show that scene-agent reasoning and adaptive fusion influence various aspects of the multi-label predictions. The A3 model yields the highest Micro-F1 score of 0.453180; the direct flow of information among the scene, target-agent, and other-agent nodes is associated with substantially improved thresholded classification performance. Meanwhile, A4 achieves the highest mAP value of 0.278457, meaning that the adaptive fusion step enhances the ordering process of the probabilities despite a slight reduction in Micro-F1 compared to A3. The distinction between the two is vital since the former assesses binary decision-making after thresholding, while the latter assesses probability orderings at various thresholds.

It must also be noted that the category-wise analysis given in Table V indicates that the benefit of contextual reasoning is not uniform across all emotion categories. The average precision for A4 is higher than that of A1 for all 26 emotion categories, yet the amount of increase varies greatly. Emotional categories like Sadness, Anger, Disconnection, and Happiness exhibit improvement; however, emotion categories with lower average precision values, like Embarrassment, Aversion, Fear, Fatigue, Sensitivity, and Yearning, prove to be more challenging.

The qualitative results present complementary support for this behavior. The successful examples in Fig. 3 demonstrate cases where the information about the scene and the agent changes the labels of predictions in a positive way, while the failed example in Fig. 4 shows that adaptive fusion can also negatively impact thresholded predictions. So, the value of the proposed method should be evaluated in terms of the overall performance on a category level.

Overall, the results indicate that through the use of the Micro-F1 score, mAP, category AP, bootstrap method, and qualitative findings, the major benefit provided by the approach is the modeling of contextual and interpersonal

relations. The contrast between A3 and A4 proves that optimizing probability ranking and thresholded classification are not the same goals.

F. Per-Category Average Precision

The AP value was calculated separately for each of the 26 categories. The AP value of A4 is greater than the value of A1 in every category listed in Table V. The AP values for A4 are greatest for Engagement, Confidence, Excitement, and Happiness; others are lower.

From the category-level table, one can see that the change in mAP is accompanied by changes at the label level, although these vary in size.

TABLE V. FINAL PER-CATEGORY AP RESULTS

No.	EMOTIC Emotion Category	A1 AP	A2 AP	A4 AP
1	Affection	0.315035	0.346010	0.338480
2	Anger	0.104340	0.094894	0.151498
3	Annoyance	0.150276	0.145350	0.154580
4	Anticipation	0.564879	0.565435	0.576285
5	Aversion	0.064305	0.064137	0.084348
6	Confidence	0.761006	0.754694	0.761472
7	Disapproval	0.127509	0.131588	0.152376
8	Disconnection	0.244374	0.262811	0.279961
9	Disquietment	0.167210	0.171058	0.176570
10	Doubt/Confusion	0.176686	0.178336	0.187716
11	Embarrassment	0.021556	0.024070	0.031498
12	Engagement	0.854545	0.864934	0.864268
13	Esteem	0.155370	0.157432	0.179174
14	Excitement	0.690954	0.699512	0.706421
15	Fatigue	0.100408	0.119506	0.140422
16	Fear	0.052422	0.057836	0.071961
17	Happiness	0.666772	0.670235	0.692305
18	Pain	0.065892	0.060228	0.088011
19	Peace	0.232054	0.233640	0.246999
20	Pleasure	0.422248	0.430982	0.423973
21	Sadness	0.199768	0.216699	0.287131
22	Sensitivity	0.063680	0.078414	0.078785
23	Suffering	0.180832	0.183180	0.219946
24	Surprise	0.075293	0.073939	0.102873
25	Sympathy	0.123645	0.138093	0.145623
26	Yearning	0.085072	0.080095	0.097214
Mean AP (mAP)		0.256390	0.261658	0.278457

G. Analysis of Difficult Emotion Categories

Category-level AP differs among emotions; lower AP means more difficult ranking by the evaluation criteria, not necessarily implying any ambiguity in the category itself.

H. Qualitative Interpretation of Scene-Agent Interaction

The qualitative analysis examines how scene-agent information affects target predictions using representative EMOTIC test samples. Fig. 3 presents three representative A3-versus-A4 cases, including two improvement cases and one strong-success case.



Fig. 3. Successful qualitative examples from the EMOTIC test set. (a) A4 improvement case (Sample 1475), (b) A4 improvement case (Sample 1195), and (c) Strong success case (Sample 667). Ground-truth multi-label annotations and predictions from the A3 scene-agent reasoning and A4 full-fusion configurations are shown for each example.

Fig. 4 shows Sample 6932, where A3 obtains $F1=0.6000$ and A4 obtains $F1=0.2222$ ($\Delta F1=-0.3778$), illustrating that fusion does not improve every individual thresholded prediction.

	Sample ID	6932				
	A3 F1	0.6000	A4 F1	0.2222	$\Delta F1$	-0.3778
	Ground truth labels	Anticipation Disconnection Engagement Excitement				
	A3 prediction	Affection Anticipation Disconnection Engagement Peace Yearning				
	A4 prediction	Affection Engagement Peace Pleasure Yearning				

Fig. 4. Failure/ambiguous qualitative example from the EMOTIC test set (Sample 6932). The A3 scene-agent reasoning configuration achieves a sample-level F1-score of 0.6000, whereas the A4 full-fusion configuration obtains 0.2222 ($\Delta F1 = -0.3778$), illustrating an individual case in which adaptive fusion produces a lower thresholded prediction score.

I. Discussion of the Main Findings

The results show three patterns: A2 changes performance only slightly relative to A1; A3 produces the largest Micro-F1 increase; and A4 improves mAP relative to A3 while reducing Micro-F1 slightly.

J. Comparison with Existing Methods

Wu et al. [3] proposed to use hierarchical context representations based on scene graphs and achieved a C-F1 score of 30.24% on EMOTIC. Chen et al. [13] introduced ICENet for the integration of facial, posture, and scene-context interactions using graph reasoning. Tran et al. [43] introduced the concept of CECNet that combined the visual data with the textual descriptions produced by the vision-language models, achieving a mAP of 43.72%. The approach presented in this work, called DICG-AF, was able to achieve a Micro-F1 score of 0.4480 and an mAP of 0.2785 with the A4 configuration. However, differences in feature extractors, input modalities, evaluation metrics, and experimental protocols limit direct comparisons. In particular, C-F1, Micro-F1, and mAP are different metrics, and the reported results do not establish a

direct performance ranking without evaluation under consistent settings (Table VI).

TABLE VI. COMPARATIVE IMPROVEMENT/DIFFERENCE ANALYSIS

Comparison	Δ Micro-F1	Δ mAP	Interpretation
A2 – A1	+0.001332	+0.005268	Small changes in both metrics
A3 – A1	+0.192468	+0.019038	Larger increases in both metrics
A4 – A1	+0.187326	+0.022067	Higher mAP; Micro-F1 remains below A3
A4 – A3	–0.005142	+0.003029	Higher mAP; slightly lower Micro-F1

VII. LIMITATIONS

The conclusions are drawn based on one benchmark, and no generalizations or causalities across datasets are made. The graph generated must be viewed relationally and not causally. In future studies, the following may be considered: automatic target detection, enhanced graph structures, adaptation of visual backbones, evaluation of cross-dataset generalizations, temporal or multimodal generalizations, calibration, and variability analysis across different training seeds.

A. Generalization, Interpretability, and Future Work

More than one emotion category can make sense at once, while unmarked categories do not have to be visually missing. The findings further show that there is behavior that depends on the specific metric used: A3 has better Micro-F1, while A4 has better mAP.

B. Multi-Label Ambiguity and Metric Behavior

The visual backbone is a fixed ImageNet-pretrained ResNet-50 that produces 2048-dimensional features. The graph structure is extracted based on visual representations and the annotations of people but not on the semantic interactions that are separately annotated; learning graph structures and end-to-end backbone adaptation remain future work.

C. Representation and Graph Construction

The evaluation employs the benchmark of the static image EMOTIC dataset and makes an assumption that target person annotations are present. The categories of audio, speech, dialogue, text, and video over time do not fall within the reported scope.

D. Dataset, Modality, and Annotation Dependence

The analysis is confined to the static image EMOTIC dataset and target person annotation. Temporal video and multimodal data fall beyond the scope of the present analysis, and incorrect localization of the target may lead to error propagation throughout the system. It will be important in the future to assess automatic target detection and temporal/multimodal frameworks.

VIII. CONCLUSION

This study introduced DICG-AF for context-aware multi-label emotion recognition in the static EMOTIC images. The pretrained ResNet-50 model extracts the scene and target agent features, and then projects them to 256 dimensions where they are modeled in a single graph containing two graph

convolutional layers. In A4, the attended target-agent representation is combined with the scene representation through adaptive gated fusion to produce the final 26-class predictions.

The final experiment used 34,318 valid person-level samples (23,705-training, 3,334-validation, 7,279-testing). Under the common test protocol, A1 obtained Micro-F1/mAP of 0.260712/0.256390; A2 0.262044/0.261658; A3 0.453180/0.275428; and A4 0.448038/0.278457. Thus, A3 provides the higher Micro-F1 while A4 provides the higher mAP, and A4 exceeds A1 in AP for all 26 categories.

The findings support the usefulness of integrated scene-agent relational reasoning within the evaluated static-image EMOTIC setting. They do not prove causal influences, automatic target detection capabilities, or generalizability beyond the evaluated dataset and modality.

REFERENCES

- [1] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context-Based Emotion Recognition Using EMOTIC Dataset," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2755–2766, Nov. 2020, doi: 10.1109/TPAMI.2019.2916866.
- [2] T. Ding, K. Zhang, S. Gao, X. Miao, and J. Xi, "A multimodal driver anger recognition method based on context-awareness," *IEEE Access*, vol. 12, pp. 118533–118550, 2024, doi: 10.1109/ACCESS.2024.3422383.
- [3] S. Wu, L. Zhou, Z. Hu, and J. Liu, "Hierarchical context-based emotion recognition with scene graphs," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 3, pp. 3725–3739, Mar. 2024, doi: 10.1109/TNNLS.2022.3196831.
- [4] N. Lu, Z. Han, and Z. Tan, "A hypergraph based contextual relationship modeling method for multimodal emotion recognition in conversation," *IEEE Trans. Multimedia*, vol. 27, pp. 2243–2255, 2025, doi: 10.1109/TMM.2024.3521738.
- [5] L. Zhang, X. Zheng, X. Chen, and L. Cui, "Three-dimensional view relationship-based context-aware emotion recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 7, pp. 13567–13578, Jul. 2025, doi: 10.1109/TNNLS.2024.3476249.
- [6] K. Fujii, D. Sugimura, and T. Hamamoto, "Hierarchical group-level emotion recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 3892–3906, 2021, doi: 10.1109/TMM.2020.3033125.
- [7] M.-H. Hoang, S.-H. Kim, H.-J. Yang, and G.-S. Lee, "Context-aware emotion recognition based on visual relationship detection," *IEEE Access*, vol. 9, pp. 90465–90474, 2021, doi: 10.1109/ACCESS.2021.3091169.
- [8] Z. Wang et al., "Visual and textual prompts in VLLMs for enhancing emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 12, pp. 12355–12368, Dec. 2025, doi: 10.1109/TCSVT.2025.3588892.
- [9] D. Li, B. Xing, and X. Liu, "Enhancing micro gesture recognition for emotion understanding via context-aware visual-text contrastive learning," *IEEE Signal Process. Lett.*, vol. 31, pp. 1309–1313, 2024, doi: 10.1109/LSP.2024.3396656.
- [10] P. Bhattacharya, R. K. Gupta, and Y. Yang, "Exploring the contextual factors affecting multimodal emotion recognition in videos," *IEEE Trans. Affective Comput.*, vol. 14, no. 2, pp. 1547–1557, Apr.–Jun. 2023, doi: 10.1109/TAFFC.2021.3071503.
- [11] C.-S. Jiang and Z.-T. Liu, "Emotion recognition via environmental context and human body," in *Proc. 2023 IEEE 6th Int. Conf. Industrial Cyber-Physical Systems (ICPS)*, Wuhan, China, 2023, pp. 1–5, doi: 10.1109/ICPS58381.2023.10128084.
- [12] N. Ayari, H. Abdelkawy, A. Chibani, and Y. Amirat, "Hybrid model-based emotion contextual recognition for cognitive assistance services," *IEEE Trans. Cybern.*, vol. 52, no. 5, pp. 3567–3576, May 2022, doi: 10.1109/TCYB.2020.3013112.

- [13] Y.-X. Chen, H.-M. Sun, C.-Y. Che, S. Feng, and R.-S. Jia, "Context-aware emotion recognition through agent-scene interactions," *Engineering Applications of Artificial Intelligence*, vol. 158, Part B, Art. no. 111581, Oct. 2025, doi: 10.1016/j.engappai.2025.111581.
- [14] W. Gong, Y. Wang, Y. Wu, S. Gao, A. V. Vasilakos, and P. Zhang, "A hybrid fusion model for group-level emotion recognition in complex scenarios," *Information Sciences*, vol. 704, Art. no. 121968, Jun. 2025, doi: 10.1016/j.ins.2025.121968.
- [15] B. A. Steward, P. Mewton, R. Palermo, and A. Dawel, "Interactions between faces and visual context in emotion perception: A meta-analysis," *Psychonomic Bulletin & Review*, vol. 32, no. 5, pp. 1987–2003, Oct. 2025, doi: 10.3758/s13423-025-02678-6.
- [16] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Trans. Affective Comput.*, vol. 14, no. 2, pp. 1236–1248, Apr.–Jun. 2023, doi: 10.1109/TAFFC.2021.3122146.
- [17] C.-S. Jiang, Z.-T. Liu, and J. She, "Hierarchical co-consistency quantization and information refining binary network for facial expression recognition in human–robot interaction," *IEEE Trans. Ind. Inform.*, vol. 20, no. 10, pp. 12178–12188, Oct. 2024, doi: 10.1109/TII.2024.3414489.
- [18] N. Perveen, D. Roy, and K. M. Chalavadi, "Facial expression recognition in videos using dynamic kernels," *IEEE Trans. Image Process.*, vol. 29, pp. 8316–8325, 2020, doi: 10.1109/TIP.2020.3011846.
- [19] Y. Huo and Y. Ge, "Multi-task emotion recognition based on dimensional model and category label," *IEEE Access*, vol. 12, pp. 75169–75179, 2024, doi: 10.1109/ACCESS.2024.3404990.
- [20] S. J. Colaco and D. S. Han, "Scalable context-based facial emotion recognition using facial landmarks and attention mechanism," *IEEE Access*, vol. 13, pp. 20778–20791, 2025, doi: 10.1109/ACCESS.2025.3534328.
- [21] J. Xu, Y. Li, G. Yang, L. He, and K. Luo, "Multiscale facial expression recognition based on dynamic global and static local attention," *IEEE Trans. Affective Comput.*, vol. 16, no. 2, pp. 683–696, Apr.–Jun. 2025, doi: 10.1109/TAFFC.2024.3458464.
- [22] S. Roy and A. Etemad, "Exploring the boundaries of semi-supervised facial expression recognition using in-distribution, out-of-distribution, and unconstrained data," *IEEE Trans. Affective Comput.*, vol. 16, no. 2, pp. 458–470, Apr.–Jun. 2025, doi: 10.1109/TAFFC.2024.3424882.
- [23] M. H. Siddiqi, I. Ahmad, Y. Alhwaiti, and F. Khan, "Facial expression recognition for healthcare monitoring systems using neural random forest," *IEEE J. Biomed. Health Inform.*, vol. 29, no. 1, pp. 30–42, Jan. 2025, doi: 10.1109/JBHI.2024.3482450.
- [24] X. Sun et al., "Design and analysis of a human–machine interaction system for researching human's dynamic emotion," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 51, no. 10, pp. 6111–6121, Oct. 2021, doi: 10.1109/TSMC.2019.2958094.
- [25] J.-H. Hsu and C.-H. Wu, "Applying segment-level attention on bi-modal transformer encoder for audio-visual emotion recognition," *IEEE Trans. Affective Comput.*, vol. 14, no. 4, pp. 3231–3243, Oct.–Dec. 2023, doi: 10.1109/TAFFC.2023.3258900.
- [26] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 985–1000, 2021, doi: 10.1109/TASLP.2021.3049898.
- [27] X. Zhang, M. Wang, X. Zeng, and X. Zhuang, "Af-CAN: Multimodal emotion recognition method based on situational attention mechanism," *IEEE Access*, vol. 13, pp. 44858–44871, 2025, doi: 10.1109/ACCESS.2024.3471613.
- [28] D. Wang, S. Liu, Q. Wang, Y. Tian, L. He, and X. Gao, "Cross-modal enhancement network for multimodal sentiment analysis," *IEEE Trans. Multimedia*, vol. 25, pp. 4909–4921, 2023, doi: 10.1109/TMM.2022.3183830.
- [29] S. Hong, H. Kang, and H. Cho, "Cross-modal dynamic transfer learning for multimodal emotion recognition," *IEEE Access*, vol. 12, pp. 14324–14333, 2024, doi: 10.1109/ACCESS.2024.3356185.
- [30] T. Wu, E. Purwanto, Y. Huang, and Y. Su, "Phy-FusionNet: A memory-augmented transformer for multimodal emotion recognition with periodicity and contextual attention," *IEEE Trans. Affective Comput.*, vol. 17, no. 1, pp. 235–246, Jan.–Mar. 2026, doi: 10.1109/TAFFC.2025.3609046.
- [31] U. Mamodiya, I. Kishor, A. Ahmed Syed, P. Sankalkar, and N. Naik, "An adaptive human–robot interaction framework using real-time emotion recognition and context-aware task planning," *IEEE Access*, vol. 13, pp. 152219–152240, 2025, doi: 10.1109/ACCESS.2025.3603738.
- [32] M. J. Dileep Kumar, M. Suresh Rao, and K. C. Narendra, "Multimodal emotion recognition: A comprehensive survey of methods, and applications," *IEEE Access*, vol. 13, pp. 201067–201097, 2025, doi: 10.1109/ACCESS.2025.3636186.
- [33] M. T. Teye, Y. M. Missah, E. Ahene, and T. Frimpong, "Evaluation of conversational agents: Understanding culture, context and environment in emotion detection," *IEEE Access*, vol. 10, pp. 24976–24984, 2022, doi: 10.1109/ACCESS.2022.3153787.
- [34] U. A. Khan, Q. Xu, Y. Liu, A. Lagstedt, A. Alamäki, and J. Kauttonen, "Exploring contactless techniques in multimodal emotion recognition: Insights into diverse applications, challenges, solutions, and prospects," *Multimedia Systems*, vol. 30, no. 2, Art. no. 115, Apr. 2024, doi: 10.1007/s00530-024-01302-2.
- [35] M. P. A. Ramaswamy and S. Palaniswamy, "Multimodal emotion recognition: A comprehensive review, trends, and challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 6, Art. no. e1563, Nov.–Dec. 2024, doi: 10.1002/widm.1563.
- [36] D. Yang, K. Yang, H. Kuang, Z. Chen, Y. Wang, and L. Zhang, "Towards context-aware emotion recognition debiasing from a causal demystification perspective via de-confounded training," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10663–10680, Dec. 2024, doi: 10.1109/TPAMI.2024.3443129.
- [37] H. Elhamedadi, S. Canavan, and P. Rosen, "AffectiveTDA: Using topological data analysis to improve analysis and explainability in affective computing," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 769–779, Jan. 2022, doi: 10.1109/TVCG.2021.3114784.
- [38] D. Al Chanti and A. Caplier, "Deep learning for spatio-temporal modeling of dynamic spontaneous emotions," *IEEE Trans. Affective Comput.*, vol. 12, no. 2, pp. 363–376, Apr.–Jun. 2021, doi: 10.1109/TAFFC.2018.2873600.
- [39] K. Zhang et al., "Real-time video emotion recognition based on reinforcement learning and domain knowledge," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1034–1047, Mar. 2022, doi: 10.1109/TCSVT.2021.3072412.
- [40] L. E. Arenas-Deseano, J. M. Ramirez-Cortes, J. Rangel-Magdaleno, and I. Cruz-Vega, "Real-time multiplatform emotion classification using CNN in a fog computing environment," *IEEE Access*, vol. 12, pp. 139988–139997, 2024, doi: 10.1109/ACCESS.2024.3450633.
- [41] L. Xiong, J. Zhang, X. Zheng, and Y. Wang, "Context Transformer and Adaptive Method with Visual Transformer for robust facial expression recognition," *Applied Sciences*, vol. 14, no. 4, Art. no. 1535, Feb. 2024, doi: 10.3390/app14041535.
- [42] E. M. G. Younis, S. Mohsen, E. H. Houssein, and O. A. S. Ibrahim, "Machine learning for human emotion recognition: A comprehensive review," *Neural Computing and Applications*, vol. 36, pp. 8901–8947, 2024, doi: 10.1007/s00521-024-09426-2.
- [43] N. Tran, T. Le, P. Ta, H. Nguyen, and H. D. Nguyen, "Contrastive context-aware emotion recognition: A multimodal framework for learning contextualized emotional representations," *Neurocomputing*, vol. 679, Art. no. 133237, May 2026, doi: 10.1016/j.neucom.2026.133237.
- [44] R. Abbas, B. Ni, R. Ma, T. Li, Y. Lu, and X. Li, "Context-based emotion recognition: A survey," *Neurocomputing*, vol. 618, Art. no. 129073, Feb. 2025, doi: 10.1016/j.neucom.2024.129073.