Pattern Recognition 165 (2025) 111640

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

SymGraphAU: Prior knowledge based symbolic graph for action unit recognition

Weicheng Xie [a,b,c], Fan Yang [a,c], Junliang Zhang [a,c], Siyang Song [d], Linlin Shen [a,c,*], Zitong Yu [e], Cheng Luo [f]

[a] School of Computer Science and Software Engineering, Shenzhen University, China
[b] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China
[c] Guangdong Provincial Key Laboratory of Intelligent Information Processing, China
[d] School of Computer Science, University of Exeter, UK
[e] Department of Computing and Information Technology, Great Bay University, China
[f] King Abdullah University of Science and Technology, Kingdom of Saudi Arabia

ARTICLE INFO

Keywords:
Action unit recognition
Symbolic logic proposition
Joint AU and expression learning
Graph convolutional network

ABSTRACT

The prior and sample-aware semantic association between facial Action Units (AUs) and expressions, which could yield insightful cues for the recognition of AUs, remains underexplored within the existing body of literature. In this paper, we introduce a novel AU recognition method to explicitly explore both AUs and Expressions, incorporating existing knowledge about their relationships. Specifically, we novelly use the Conjunctive Normal Form (CNF) in propositional logic to express these knowledges. Thanks to the flexible and explainable logic proposition, our method can dynamically build a knowledge base specifically for each sample, which is not limited to fixed prior knowledge pattern. Furthermore, a new regularization mechanism is introduced to learn the predefined rules of logical knowledge based on embedding graph convolutional networks. Extensive experiments show that our approach can outperform current state-of-the-art AU recognition methods on the BP4D and DISFA datasets. Our codes will be made publicly available.

1. Introduction

Facial expression analysis occupies a pivotal role in human communication, human–machine interface and psychology. Traditional facial expression recognition systems assign discrete emotional states to rudimentary categories (*e.g.,* happy, angry, surprise, fear, sad, contempt, disgust and neutral [1]). However, these categorical expressions fail to accurately describe all possible human facial emotion. As a result, a more objective and comprehensive expression system, Facial Action Coding System (FACS) [2], has garnered substantial attention [3,4] for its application potential in robot expression generation, expression parsing and understanding, micro-expression recognition, etc. FACS decomposes facial expressions into individual components of muscle movements, termed Action Units (AUs, *e.g.,* AU4 Brow Lowerer, AU6 Cheek Raiser, etc.).

Consequently, quite a few approaches for AU recognition have surfaced within the last half-decade. These works all treat facial AU recognition as a multi-label classification problem since multiple AUs can be activated simultaneously. Notably, some of them have endeavored to model the interplay among AUs, such as AU dependencies [6,7] and spatial interrelationships [8]. Other works also introduced supplementary cues (*e.g.,* facial landmarks [9], textual descriptions [10] and face synthesis [11]) germane to AUs, with the aim of benefiting AU recognition tasks. However, the intuitive correlations between AUs, which embody local/fine-grained emotion cues, and facial expressions, as well as convey global/coarse-grained emotional signals [12], remain a terrain largely underexplored. These two levels of facial emotion descriptors engage in mutual interaction and can enhance recognition performances and interpretability of both sides. While previous works [12–14] have predominantly revolved around elevating expression recognition performance by AU–Expression correlations where AUs serve as auxiliary roles, very few attentions have been drawn to the impact of AU–Expression correlations on AU recognition. As shown in Fig. 1, some subtle action units cannot be accurately recognized due to the absence of global emotional cues (expression). The existing work [5] also explored AU–Expression correlations as generic and fixed knowledge, while it disregarded the variability inherent in AU–Expression pairs within dynamic contexts.

Therefore, we argue that the broader contextual information conveyed by facial expressions can serve as effective informative cues

* Corresponding author at: School of Computer Science and Software Engineering, Shenzhen University, China.
E-mail address: llshen@szu.edu.cn (L. Shen).

https://doi.org/10.1016/j.patcog.2025.111640
Received 5 June 2024; Received in revised form 19 March 2025; Accepted 24 March 2025
Available online 2 April 2025

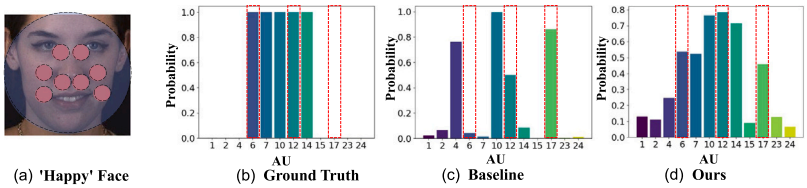

Fig. 1. (a) A sample with AU and expression labels, red circles represent the specific positions of each AU on the face. The abscissa of the bar chart represents different AUs. As mentioned in [2,5], AU6 (Cheek Raiser) and AU12 (Lip Corner Puller) usually appear together with Happy; AU12 and AU17 (Chin Raiser) are controlled by conflicting facial muscles that often cannot occur simultaneously. The red dotted box represents the AU recognition performances of different similarities between the ground truth and the results by our method via explicitly introducing the regularization of this prior knowledge and the baseline. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

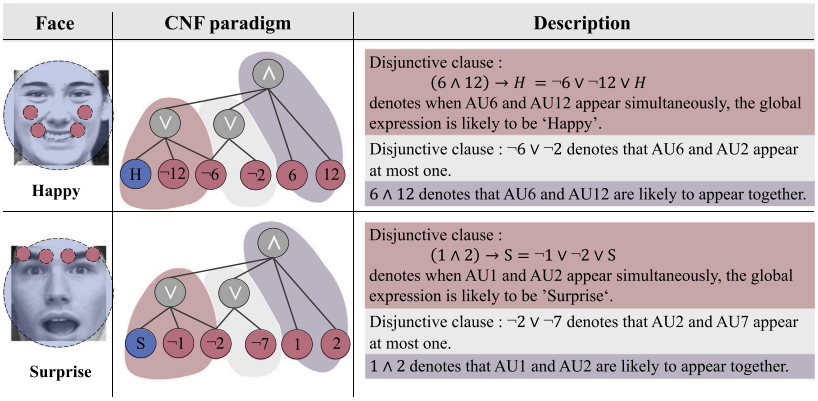

Fig. 2. Illustration of explicit AU–AU and AU–Expression knowledge construction through our symbolic graph. For each face display, we can build *specific* (i) AU–Expression (in red background); (ii) AU–AU co-occurrence (in purple background); and (iii) AU–AU mutual exclusion (in gray background) relationships by CNF. 'H' denote 'Happy', '1' denotes 'AU1', '∧' and '∨' denote 'AND' and 'OR', respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for detecting AUs, as FACS [2] elucidates strong semantic correlations between AUs and basic facial expressions [13]. A primary limitation of prevalent methodologies [3,8] for modeling AU relationships lies in their neglect of the AU–Expression interplay that can provide prior knowledge for AUs **(Problem 1)**. Second, the relationships among AUs they endeavor to model often lack a flexibility and interpretative framework **(Problem 2)**. While Cui et al. [5] has introduced a constraint optimization that leverages Bayesian Networks (BN) to encode generic knowledge pertaining to AU–Expression probabilistic dependencies, it relies upon predefined and static prior knowledge, rendering it to lose the flexibility to construct relationship tailored to the facial characteristics of an unique sample **(Problem 3)**.

This paper introduces a Symbolic Graph-based AU recognition (SymGraphAU) paradigm, which establishes a neural-symbolic framework and uses flexible symbolic prior knowledge to introduce AU recognition for the first time. Within this framework, we offer three distinct solutions to address the aforementioned challenges. Firstly, a strategy termed Joint Tasks for Node Feature Learning (JFL) is proposed, which enabled the backbone to simultaneously capture global expression and local AUs patterns **(Addressing problem 1)**. Secondly, we employ the flexible CNF paradigm to generate unique and precise knowledge for each individual. In this way, we can mine unique knowledge adapt to each dataset, transcending the usage of generic and static prior knowledge **(Addressing problem 3)**. As depicted in Fig. 2, clauses in a CNF paradigm clearly express statements from AU–Expression correlations to AU–AU interactions. Thirdly, we present a prior Expression Knowledge Regularized AU recognition (EKR), explicitly formulating prior knowledge relating to AU–AU pairs and AU–Expression correlations through CNF paradigms, thus making our framework more interpretable in contrast to prevailing AU relation learning paradigms **(Addressing problems 1 & 2)**. Our work's contributions are summarized as:

- We propose a novel AU recognition method which builds prior knowledge between AU and expressions through CNF. To the best of our knowledge, this is the first work that explicitly models the relationship between AU–AU and AU–Expression via logic rules, which can dynamically construct prior knowledge to recognize AUs.
- We design a regularization mechanism to enable the AU recognition to learn the prior knowledge of logic rules from expressions in terms of graph convolutional networks, which can explore the knowledge of AU–AU and AU–Expression to assist AU recognition.
- Extensive experimental results on two mainstream datasets, *i.e.,* BP4D and DISFA, demonstrate the effectiveness of our symbolic graph-based method in terms of F1-score.

2. Related works

2.1. AU recognition

Due to the progresses in deep learning, current state-of-the-art (SOTA) AU recognition methods mainly resort to deep models. Eleftheriadis et al. [15] applied the convolutional neural network to AU recognition, which provided a reliable face representation for this task. Zhao et al. [16] treated the AU recognition as a multi-label recognition task, and simultaneously captured the key region and the AU dependencies. Additional studies [3,6] represent AUs as a graph where each AU is treated as a node, and use the connection of nodes in the graph structure to describe the relationship between AUs. Luo et al. [8] consider more complex relationship that may exist between AU, and propose a multi-dimensional edge to better describe this relationship. These methods only consider the AU itself or the relationship between AUs.

However, AUs are often subtle, it is thus difficult to simply learn AU or learn the relationship between AUs. As another signal for describing human emotions, expression is more global and has a strong correlation with AU [2]. Zhang et al. [17] propose to jointly train AU classifiers

using general knowledge such as probabilities over AUs rather than annotations of AUs. Cui et al. [5] use Bayesian Networks to capture AU–Expression dependencies for joint expression and AU recognition, while they only use general knowledge and ignore the specificities shown by different people. Our method uses a flexible CNF paradigm to construct a specific knowledge representation for each sample.

2.2. Neural-symbolic system

Neural-symbolic systems have the combined advantages of both neural systems (powerful learning capacity and superior perception intelligence) and symbolic systems (exceptional cognitive intelligence) [18,19]. Neural symbolic systems, *e.g.,* [19] can be mainly divided into the categories of learning for reasoning [20], reasoning for learning [18], and learning–reasoning interaction [21]. The learning for reasoning incorporates the strengths of neural networks to facilitate finding solutions of symbolic systems. The reasoning for learning uses symbolic systems to support neural network learning. The learning–reasoning interaction proposes to learn the symbol systems and the neural network in a collaborative manner.

Recently, a few works have focused on using symbolic systems to regularize the training of neural networks. Diligenti et al. [22] encoded logic formulas (propositional logic or first-order logic) into real-valued functions as regularization terms for neural models. Xu et al. [23] enabled neural networks to use the reasoning capabilities of propositional logic to improve their learning capabilities. Xie et al. [18] used explicit knowledge d-DNNF to enhance the ability of network relationship prediction.

However, it remains a challenging open problem to design a symbolic learning system in a flexible and efficient manner. Especially, to the best of our knowledge, there is no work integrating explicit and sample-aware propositional logic into AU recognition. Previous works [5,8] implicitly model prior knowledge via feature representations, while the interpretability and flexibility are unexplored. In this work, we design regularized models to integrate symbolic knowledge into the network training process, so as to encourage networks to adhere to the symbolic knowledge during training.

3. Methodology

The overview of our approach is depicted in Fig. 3.

Preliminary. This work is inspired by the propositional logic, where a proposition is a statement which is either True or False. A formula is a compound of propositions connected by logical connectives, *e.g.,* $\neg, \wedge, \vee, \Rightarrow$. We use the CNF to explicitly express the binary relationships between AUs and the relationship between AUs and expressions, which is the conjunction of a series of disjunctive clauses. Let U be the set of propositional variables. A sentence in CNF is defined as a rooted undirected acyclic graph where each leaf node is labeled with u or $\neg u$, $u \in U$; Each internal node is labeled with $\wedge$ or $\vee$ and can have arbitrary number of child nodes.

3.1. Joint tasks for node feature learning

To initially encode global expression information, we use decoupled recognition heads to train backbone networks from different granularities (*e.g.,* fine-grained AUs and coarse-grained expressions). Furthermore, expression and AU embedding are obtained for subsequent logic training, differing from other methods [18] using pre-defined node semantics (word embedding).

For the training of the expression branch, it should be noticed that the basic expressions are not thoroughly annotated in current facial AU datasets. Fortunately, according to FACS [2], an expression has primary and secondary relationships with AUs, we thus leverage these relationships between expressions and AUs to predict the expression labels, where the relationships are summarized in the supplementary materials. Specifically, we build a matrix $\mathbf{M}_{A-E} \in \mathbb{R}^{N_\mathrm{a} \times N_\mathrm{e}}$ (N_a and N_e denote the numbers of AUs and expressions, respectively), mapping AU annotations to expression labels. Each item $\mathbf{M}_{A-E}(i,j)$ in the matrix is a value conditioned on AU–Expression relationship, *i.e.,* the i_{th} AU and the j_{th} expression: (1) $\mathbf{M}_{A-E}(i,j) = 1-\beta$, if the AU–Expression pair has primary relevance; (2) $\mathbf{M}_{A-E}(i,j) = 0.5$, if the pair has secondary relevance; (3) $\mathbf{M}_{A-E}(i,j) = \beta$, if the pair has no relevance, where $\beta = 0.1$ is a hyperparameter that will be analyzed in the experimental section. As a result, given an AU annotation $Y^a = [y^a_1, \ldots, y^a_{N_\mathrm{a}}] \in \mathbb{R}^{1\times N_\mathrm{a}}$, we can estimate the expression label as:

$$k^e = \mathrm{argmax}(Y^a \mathbf{M}_{A-E}) \tag{1}$$

In instances where none of the AUs exhibit activation, the corresponding expression is annotated as 'Neutral'. It is worthwhile to note that the associated expression is also designated as 'Neutral' in a small number of cases when there is a little AU activation. Consequently, to make the generation process of the expression pseudo-label unified, we ignore the impact of these small number of samples on model training. And the label vector is indicated by $Y^e = [y^e_1, \ldots, y^e_{N_\mathrm{e}}] \in \mathbb{R}^{1\times N_\mathrm{e}}$ as:

$$y^e_j = \begin{cases} 1, & j = k^e \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where k^e is the pseudo expression label formulated in Eq. (1) and $1 \le j \le N_\mathrm{e}$.

Given an input image I, we first produce both AU and expression representations (V^a and V^e) for further logic graph construction. The feature extractor consists of a backbone $\mathcal{F}(\cdot)$ and two decoupled heads, where one is followed by an AU detector $\mathcal{H}_a(\cdot)$ and the other one is added by an expression classification layer $\mathcal{H}_e(\cdot)$. Specifically, the backbone projects a face image I into a feature map implying primary cues of this face. Then N_a AU-specific and N_e expression-specific feature extractors are further processed to obtain AU representations $V^a = [V^a_1, \ldots, V^a_{N_\mathrm{a}}]$, AU prediction probabilities $p^a = \{p^a_1, \ldots, p^a_{N_\mathrm{a}}\}$, expression representations $V^e = [V^e_1, \ldots, V^e_{N_\mathrm{e}}]$ and expression prediction probabilities $p^e = \{p^e_1, \ldots, p^e_{N_\mathrm{e}}\}$. Each extractor (*i.e.,* AU extractor, expression extractor) contains a Sigmoid function, a fully connected layer (FC) and a global average pooling (GAP) layer.

Then, to make these representations contain meaningful semantics, we propose to supervise them via joint tasks. Specifically, a shared AU detector $\mathcal{H}_a$ predicts N_a AU occurrence probabilities $p^a = \{p^a_1, \ldots, p^a_{N_\mathrm{a}}\}$, which is supervised by a weighted asymmetric loss [8]:

$$\mathcal{L}_\mathrm{wa} = -\frac{1}{N_\mathrm{a}} \sum_{i=1}^{N_\mathrm{a}} w_i [y^a_i \log(p^a_i) + (1-y^a_i) p^a_i \log(1-p^a_i)] \tag{3}$$

where y^a_i and w_i are the ground truth and the weight for the i_{th} AU, respectively. Normally, the weight $w_i = N_\mathrm{a}(1/r_i)/\Sigma_{k=1}^{N_\mathrm{a}}(1/r_k)$ is computed according to the occurrence rate r_i of the i_{th} AU specific to the training set. This weight strategy has been proven in [8] to be effective in AU recognition for the case of imbalanced samples. Meanwhile, the basic expression branch is also supervised by a multi-label Softmax loss, which predicts N_e expression occurrence probabilities $p^e = \{p^e_1, \ldots, p^e_{N_\mathrm{e}}\}$ according to the loss:

$$\mathcal{L}_\mathrm{we} = -\frac{1}{N_\mathrm{e}} \sum_{j=1}^{N_\mathrm{e}} [y^e_j \log(p^e_j) + (1-y^e_j) p^e_j \log(1-p^e_j)] \tag{4}$$

where y^e_j and p^e_j are the ground truth and predicted probabilities for the j_{th} expression, respectively.

3.2. Graph construction of logic knowledge embedding

In this section, we first introduce the relationship between AUs and expressions, as well as the binary relationships between AUs. Then, we introduce how to use symbolic knowledge (*e.g.*, CNF) to express this knowledge. It is worth noting that the propositional logic we use is entirely based on the knowledge of predefined AUs and expressions.

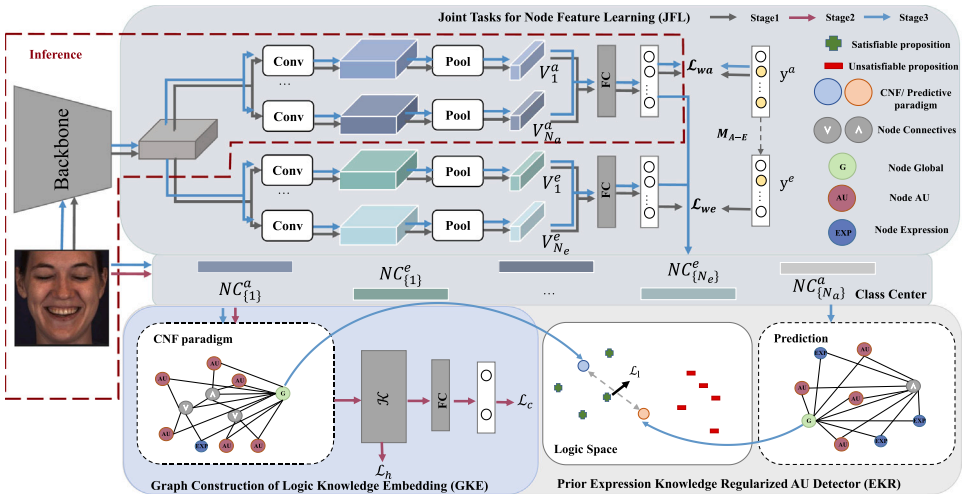

Fig. 3. Pipeline of the proposed approach of AU–Expression relationship modeling. The whole process is divided into three stages. The first stage uses the AU Ground Truth and general prior knowledge to jointly train the AU detector and the expression classifier. The second stage generates specific CNF paradigms and random satisfiable and unsatisfiable propositions for each sample, which are used to train a prior knowledge embedder. In the last stage, we use the knowledge embedder to regularize the AU detector. $\mathcal{K}$ denotes a Symbolic Knowledge Embedder with a multi-layer GCN.

In this section, we first introduce the relationship between AUs and expressions, as well as the binary relationships between AUs. Then, we introduce how to use symbolic knowledge (*e.g.*, CNF) to express this knowledge.

3.2.1. Prior expression knowledge

We construct a storing pre-defined rules, including two kinds of complementary prior knowledge as follows:

(i) AU–Expression inference. An expression will more probably appear if more relevant (primarily and secondarily relevant) AUs are activated. For example, if AU6 (Cheek Raiser) and AU12 (Lip Corner Puller) simultaneously appear on a face, we can infer that the corresponding expression is likely to be 'Happy'. We call this inference from AUs to expression as AU–Expression implications, and formulate the above example of AU–Expression relation as a logical rule of $(\text{AU6} \wedge \text{AU12}) \Rightarrow \text{Happy}$. This logic rule can be also formulated in CNF as $\neg\text{AU6} \vee \neg\text{AU12} \vee \text{Happy}$:

$$\begin{aligned}(\text{AU6} \wedge \text{AU12}) \Rightarrow \text{Happy} &= \neg(\text{AU6} \wedge \text{AU12}) \vee \text{Happy} \\ &= \neg\text{AU6} \vee \neg\text{AU12} \vee \text{Happy}\end{aligned} \tag{5}$$

(ii) AU dependency. AUs are a set of facial muscle movements, which mutually influence each other. For example, AU1 and AU2 are usually activated simultaneously as they are all innervated by the frontalis muscle. This positive dependency between two AUs is called 'Co-occurrence', and we describe it as a disjunctive clause (the 1st row of Eq. (6)):

$$\begin{cases} \text{AUm} \wedge \text{AUn}, & \text{AUm and AUn are Co-occurrence} \\ \neg\text{AUm} \vee \neg\text{AUn}, & \text{AUm and AUn are Mutual exclusion} \end{cases} \tag{6}$$

In contrast, AU2 (Outer Brow Raiser), controlled by the Frontalis, Pars Lateralis muscle, and AU6 (Cheek Raiser), controlled by the Depressor Orbicularis Oculi, Pars Orbitalis muscle, are exclusive since they rarely appear together. We call this negative AU dependency as 'Mutual Exclusion' as the 2nd row of Eq. (6). AU pairs with 'Co-occurrence' or 'Mutual Exclusion' dependencies are summarized in the supplementary material.

3.2.2. Logic graph

After defining symbolic knowledge in CNF, we leverage a Graph Convolutional Network (GCN) [24] to represent this logical knowledge. A logic formula can be represented as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with N_n nodes $\mathcal{V} = \{z_b\}$, and edges $\mathcal{E} = \{(z_b, z_c)\}$. Individual nodes are either propositions (leaf nodes) or logical operators ($\wedge$ and $\vee$).

The constructed logic graph is data-specific and assigned as the graph topology of a multi-layer GCN (*i.e.,* the adjacency matrix $A \in \mathbb{R}^{N_\text{n} \times N_\text{n}}$ is defined by the logic graph with $A(b, c) = 1$, if nodes b and c are connected in the logic graph). Then, through propagation of the GCN layer, we can get node features $Q^{(l+1)}$ from the $(l+1)_{th}$ GCN layer:

$$Q^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} Q^{(l)} W^{(l)}) \tag{7}$$

where $\sigma(\cdot)$ is a non-linear activation; $\tilde{A} = A + E$ ($\tilde{A} \in \mathbb{R}^{N_\text{n} \times N_\text{n}}$) denotes the summation of the adjacency matrix A with added self-connections $E \in \mathbb{R}^{N_\text{n} \times N_\text{n}}$; $\tilde{D} \in \mathbb{R}^{N_\text{n} \times N_\text{n}}$ is a diagonal degree matrix with $\tilde{D}(b, b) = \sum_c \tilde{A}(b, c)$; and $W^{(l)} \in \mathbb{R}^{ind^l \times ond^l}$ is a learnable weight matrix, ind^l and ond^l denote input and output dimensions of the l_{th} layer graph convolution, respectively.

For obtaining the node representation of our logic graph, since fine-grained AUs and expression lack suitable predefined representations, we could not use pre-trained word vectors as leaf nodes [18]. Thus, we alternatively resort to the exponential moving average strategy, *i.e.,* assigning the feature of each leaf node (*i.e.,* one of N_a AUs or N_e Expressions) in terms of its class center as:

$$nC^t = \begin{cases} \alpha \cdot nC^{t-1} + (1-\alpha) R^t, & t > 1 \\ R^t, & t = 1 \end{cases} \tag{8}$$

where $R^t \in \mathbb{R}^{1 \times nd}$ denotes a certain AU or expression representation, *i.e.*, V_i^a or V_j^e of the t_{th} sample, $1 \le t \le N_\text{s}$, N_s is the number of samples in the training set; α is the momentum parameter and set as 0.1. Based on Eq. (8), we obtain $NC^a_{[i]}$ and $NC^e_{[j]}$, *i.e.,* the class centers of i_{th} AU and j_{th} expression, $1 \le i \le N_\text{a}$ and $1 \le j \le N_\text{e}$. It should be noted that we obtain the class center of each AU and expression through Eq. (8) after JFL and fix it in the latter stages.

Different from the leaf nodes $\{NC^a_{[i]}, 1 \le i \le N_\text{a}\}$ and $\{NC^e_{[j]}, 1 \le j \le N_\text{e}\}$ obtained by Eq. (8), AND and OR nodes are kinds of parameters need to be learned, represented by $NC^{AND}, NC^{OR} \in \mathbb{R}^{1 \times nd}$, respectively. Due to their semantic differences, we use the orthogonal random initialization of these two representations. Finally, due to the uncertainty of the number of nodes in the knowledge graph, as shown in Fig. 4, we use the global node $NC^{Global} \in \mathbb{R}^{1 \times nd}$ to connect all nodes in the graph and represent the entire logic graph. It is obtained by stacking a ReLU function and a global average pooling (GAP) layer after the backbone of $\mathcal{F}(\cdot)$.

Consequently, there are four different types of nodes in our logic graph (*e.g.,* leaf nodes, two logical operators and one global node). Unlike previous work [18] that has to customize a W^l to learn each type of

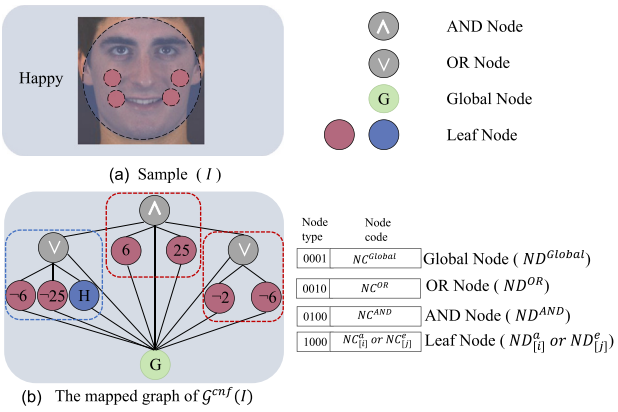

Fig. 4. Illustration of the encoding process of a logic diagram. (a): The AU label and expression pseudo-label of an input sample I. (b): CNF knowledge graph generated based on sample label knowledge. The red dotted box represents the binary relationship between AUs in Eq. (6). The blue dotted box represents the AU–Expression relationship in Eq. (5). Each node representation, *e.g.,* ND^{Global}, contains an one-hot vector denoting the node type, *e.g.,* 0001, and a vector representing the semantics of each node, *e.g.,* NC^{Global}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

node representation in Eq. (7), we resort to a resource-friendly manner, *i.e.,* using four-bit binary codes (0001, 0010, 0100, and 1000) as prefixes and concatenating designed prefixes to their node representations to distinguish their types.

Finally, as illustrated in Fig. 4, we get the representations for four kinds of nodes, *i.e.,* global node ($ND^{Global} \in \mathbb{R}^{1\times(nd+4)}$), two logical operator nodes ($ND^{AND}, ND^{OR} \in \mathbb{R}^{1\times(nd+4)}$) and leaf node ($ND^{a}_{[i]}$ for the i_{th} AU, $ND^{e}_{[j]}$ for the j_{th} expression $\in \mathbb{R}^{1\times(nd+4)}$). Meanwhile, we use the same W^l in Eq. (7) to learn the representation mappings of different nodes.

3.3. Prior expression knowledge regularized AU detection

In this section, we introduce how to regularize the training of the AU detector with a GCN embedded with the prior logical knowledge.

For every input image, we first generate the CNF paradigm for each sample, as shown in Fig. 4. Since there is a certain prediction deviation in the pseudo-labels of expressions, the relationship of AU–Expression may be uncertain. Therefore, we add the AU–Expression relationship to the CNF paradigm with a large uncertain probability p^{uct} rather than using these CNF definitively (the selection and impact of p^{uct} on algorithm performance will be discussed in section 4.2 of the supplementary materials). In this way, disjunctive clauses of AU–Expression relations can be flexibly added to the CNF paradigm in the form of reducing the over-learning possibility. For the binary relationship between AUs, we comprehensively consider prior knowledge and sample true labels. Specifically, we incorporate the real co-occurrence relationships of samples, rather than only predefined knowledge, into the CNF paradigm. Meanwhile, only mutually exclusive relationships that exist in both the true label of the sample and the predefined knowledge will be added to the CNF paradigm. In this way, we can model binary relationships between AUs more accurately.

Then, Pysat [25] is employed to generate the satisfiable and unsatisfiable propositions for each CNF paradigm (*e.g.,* $\mathcal{G}^{cnf}(I)$ in Fig. 4), where this paradigm is employed as the input of Pysat. For the output of Pysat, each leaf node, *e.g.,* an AU or expression, is assigned a Boolean value, to assess the satisfiability ('True' is returned) or unsatisfiability ('False' is returned) of the CNF paradigm. Based on this, we generate satisfiable and unsatisfiable graph representation, *i.e.,* $\mathcal{G}^{s}(I)$ and $\mathcal{G}^{us}(I)$ as follows:

$$\begin{cases} \mathcal{G}^{s}(I) = \bigwedge_{i=1}^{N_a} q^{a}_{i,s} ND^{a}_{[i]} \bigwedge_{j=1}^{N_e} q^{e}_{j,s} ND^{e}_{[j]} \wedge ND^{Global} \\ \mathcal{G}^{us}(I) = \bigwedge_{i=1}^{N_a} q^{a}_{i,us} ND^{a}_{[i]} \bigwedge_{j=1}^{N_e} q^{e}_{j,us} ND^{e}_{[j]} \wedge ND^{Global} \end{cases} \tag{9}$$

where $q^{a}_{i,s}$ and $q^{e}_{j,s}$ (obtained by Pysat) denote the assigned values of the i_{th} AU and j_{th} expression when the CNF paradigm, *i.e.,* $\mathcal{G}^{cnf}(I)$ is satisfiable (or a 'True' value is returned for $\mathcal{G}^{cnf}(I)$); $q^{a}_{i,us}$ and $q^{e}_{j,us}$ denote the corresponding values when the CNF paradigm *i.e.,* $\mathcal{G}^{cnf}(I)$ is unsatisfiable. $ND^{a}_{[i]}, ND^{e}_{[j]}, ND^{Global}$ are the node representations defined in Fig. 4.

To fine-tune the AU detector to enable networks to better learn prior knowledge using logic embedder, we align the CNF paradigm and prediction graph in the logic space. Specifically, we first build the prediction distribution into a proposition graph during fine-tuning, which is formulated as:

$$\mathcal{G}^{p}(I) = \bigwedge_{i=1}^{N_a} p^{a}_{i} ND^{a}_{[i]} \bigwedge_{j=1}^{N_e} p^{e}_{j} ND^{e}_{[j]} \wedge ND^{Global} \tag{10}$$

where p^{a}_{i} and p^{e}_{j} are the predicted probabilities of the i_{th} AU and the j_{th} expression, respectively. Eq. (10) means connecting all predicted AU, expression and a global node via $\wedge$ to get the prediction graph of $\mathcal{G}^{p}(I)$.

Based on the graph representations of $\mathcal{G}^{cnf}(I)$ in Fig. 4, $\mathcal{G}^{s}(I)$, $\mathcal{G}^{us}(I)$ in Eq. (9), and $\mathcal{G}^{p}(I)$ in Eq. (10), we map them to the logic space by a knowledge embedder, representing the prior knowledge by a two-layer GCN:

$$\begin{aligned} \mathcal{Q}(X) &= \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}Q^{(2)}W^{(2)}) \\ &= \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}\sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}XW^{(1)})W^{(2)}) \end{aligned} \tag{11}$$

where X denotes one of the graph representations of $\{\mathcal{G}^{cnf}(I), \mathcal{G}^{s}(I), \mathcal{G}^{us}(I), \mathcal{G}^{p}(I)\}$.

Based on the logic embedder $\mathcal{Q}(\cdot)$ in Eq. (11), we align the prediction graph and the knowledge graph in the logic space via a regularization term, *i.e.,* aligning the output probability with the predefined prior knowledge as:

$$\mathcal{L}_1 = \sum_{t=1}^{N_s} \| S(\mathcal{Q}(\mathcal{G}^{cnf}(I^{(t)}))) - S(\mathcal{Q}(\mathcal{G}^{p}(I^{(t)}))) \|_2^2 \tag{12}$$

where N_s denotes the number of samples; $S(\cdot)$ denotes the operation of selecting the global node; $\mathcal{G}^{cnf}(I^{(t)})$ denotes CNF graph of the t_{th} sample $I^{(t)}$, which is constructed from knowledge as in Fig. 4; $\mathcal{G}^{p}(I^{(t)})$ denotes a proposition graph constructed from AU prediction and formulated in Eq. (10).

Minimizing $\mathcal{L}_1$ in Eq. (12) actually encourages the model to adhere to our predefined prior knowledge during the fine-tuning process. Since aligning the distributions of these two graphs implicitly aligns the distributions of the prediction and the satisfiable proposition, the trained network can learn the rules of prior knowledge.

3.4. Training strategy

In this section, we introduce the three-stage training scheme to optimize our model, including the corresponding modules of JFL, GEK and EKR.

In the first stage, we train the backbone and the detector of JFL, aiming to obtain the semantic representations of AU and expression while combining their cues with different granularities during training. Due to the obvious category imbalance in the dataset, we use the weighted asymmetric loss in Eq. (3). For expression recognition module, we use the multi-label Softmax loss formulated in Eq. (4), rather than the traditional cross-entropy loss for training (an ablation study of these losses is appended in the supplementary materials). The total loss in this stage is formulated as:

$$\mathcal{L}_{jf} = \mathcal{L}_{wa} + \gamma \mathcal{L}_{we} \tag{13}$$

where γ denotes a regularization coefficient.

In the second stage, we use graph networks to learn a knowledge embedder and train AND–OR node representations using a true–false proposition classifier. A triplet loss for hard sample mining is used to

align the logistic distributions, which is formulated as:

$$\mathcal{L}_{\mathrm{h}} = \sum_{t=1}^{N_s} \|\mathcal{S}(\mathcal{Q}(\mathcal{G}^{cnf}(I^{(t)}))) - \mathcal{S}(\mathcal{Q}(\mathcal{G}^{s}(I^{(t)})))\|_2^2 - \sum_{t=1}^{N_s} 1_{d_s > d_{us}} \|\mathcal{S}(\mathcal{Q}(\mathcal{G}^{cnf}(I^{(t)}))) - \mathcal{S}(\mathcal{Q}(\mathcal{G}^{us}(I^{(t)})))\|_2^2 \tag{14}$$

where $\mathcal{G}^{s}(I^{(t)})$ and $\mathcal{G}^{us}(I^{(t)})$ denote the satisfiable and unsatisfiable propositions of $I^{(t)}$ formulated in Eq. (9); d_s or d_{us} denotes the Euclidean distance between the satisfiable propositions, *i.e.,* $\mathcal{S}(\mathcal{Q}(\mathcal{G}^{s}(I^{(t)})))$ or the unsatisfiable proposition, *i.e.,* $\mathcal{S}(\mathcal{Q}(\mathcal{G}^{us}(I^{(t)})))$, and CNF, *i.e.,* $\mathcal{S}(\mathcal{Q}(\mathcal{G}^{cnf}(I^{(t)})))$.

Furthermore, we use a classifier to classify satisfiable and unsatisfiable propositions and help learning prior logical knowledge and AND–OR semantic nodes. Specifically, we frame it as a binary classification task with the loss:

$$\mathcal{L}_{\mathrm{c}} = y^{logic} log(p^{logic}) + (1 - y^{logic}) log(1 - p^{logic}) \tag{15}$$

where p^{logic} and y^{logic} are the prediction and ground truth of logic graph (*i.e.,* satisfiable or unsatisfiable). Total loss in this stage is formulated as:

$$\mathcal{L}_{\mathrm{le}} = \mathcal{L}_{\mathrm{h}} + \lambda \mathcal{L}_{\mathrm{c}} \tag{16}$$

where λ is a regularization coefficient. In order to avoid the logic embedder overfitting to fixed satisfiable/unsatisfiable propositions that are harmful to the embedding of logical knowledge, we randomly generate satisfiable and unsatisfiable propositions during the training process instead of fixing them.

In the third stage, we use a logic loss in Eq. (12) to align the distributions of knowledge and predicted outcomes in the logic space. The total loss is formulated as:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{wa}} + \mu \mathcal{L}_{\mathrm{l}} \tag{17}$$

where $\mathcal{L}_{\mathrm{wa}}$ denotes weighted asymmetric loss formulated in Eq. (3) and μ is a regularization coefficient. For clarity, we present the flow of our algorithm in the supplementary materials.

For the inference, our algorithm only needs to use the basic AU detector without forward propagation of the graph network, i.e., does not require additional overhead compared with other graph-based methods [3,8].

4. Experiments

4.1. Datasets and implementation details

We evaluate the performance of our approach on two widely-used datasets: BP4D [26] and DISFA [27]. BP4D contains a total of 146,847 face images with labeled AUs from 41 young adult subjects (23 females and 18 males). DISFA contains 130,815 frames of images of 27 subjects (12 females and 15 males) of different races, and each frame is annotated with the occurrence labels of multiple AUs.

For both datasets, we use MTCNN [28] to perform face detection and alignment for each frame and crop it to 224×224 as the input. We follow the same protocol as previous studies [6,8] to conduct subject-independent cross-validation with three folds for each dataset, and report the average results over these folds.

We use the common metric following previous works [6,8], *i.e.,* frame-based F1-score, to evaluate the performance of our approach, which is formulated as $F1 = \frac{2Precision \cdot Recall}{Precision + Recall}$. *Precision* is the percentage of all predicted positive samples that are actually positive. *Recall* is the probability of being predicted to be a positive sample among the actual positive samples. We report the F1-score for each AU and the averaged F1-score over AUs following studies [6,8]. During the training, we use ResNet50 or Swin Transformer (pretrained on ImageNet1k as the backbone), and employ the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a weight decay of 5e−4. We set γ in Eq. (13), λ in Eq. (16) and μ in Eq. (17) to 0.05, 0.1 and 0.1, respectively. We train the proposed model for 45 epochs, including 20 epochs at the first stage (the initial learning rate is 1e−4), 5 epochs at the second stage (the initial learning rate is 1e−2) and 20 epochs at the third stage (the initial learning rate is 1e−6), with a batch size of 32. All the experiments are conducted using NVIDIA P100 GPUs based on the open-source PyTorch platform.

4.2. Comparison with the state of the arts

In this section, we compare our method with several SOTA methods on both datasets, where works [3,6–8,41] based on simple graphs of AU relationships are also compared. To make the comparison fair, we only compare our approach with static face-based methods that did not remove any frames from the dataset or employ additional training data.

Table 1 reports the occurrence recognition results of 12 AUs on BP4D. It can be seen that the proposed SymGraphAU allows two backbone networks (ResNet-50 and Swin TransformerBase (Swin-B)) to achieve an overall F1 score no lower than the other listed methods. And the result of CNN base is improved by an average of 0.5% compared to the state of the arts (FAN-Trans [35]). Specifically, our method achieves the top-three performances for 6 out of 12 AUs' recognition (*e.g.,* AU6, AU7, AU12, AU15, AU17 and AU23). According to Table 2, our method helps the learned network to achieve the SOTA performance in terms of the mean F1-score. Meanwhile, our transformer-based method outperforms the other listed methods, with a 0.7% average improvement over the state of the arts (FAN-Trans [35]).

4.3. Algorithm analysis

A1. Ablation Study of Different Modules. To evaluate the influence of each module in our pipeline, Table 3 presents the average AU recognition results of different variants. It shows that our JFL can well capture different granularities of facial knowledge (global expression information and fine-grained AU information), which helps networks learn robust AU representation. JFL achieved performance improvements of 1.3%, 0.4% on BP4D and improvements of 0.9%, 0.8% on DISFA, respectively. In addition, our EKR uses the proposed prior knowledge regularization to encourage the model to obey the AU binary relationship and the AU–Expression relationship, and achieved the improvements of 0.8%, 0.3% on BP4D and 0.9%, 0.3% on DISFA, respectively.

A2. Sample Analysis. In this part, we analyze the performance of our algorithm with two examples from BP4D and DISFA, and showcase the prediction results under different settings for them in Fig. 5. When the coarse-grained emotional cues and prior knowledge regularization are not used, Fig. 5 shows that the baseline model yields false and missing activations, while these activations can be well reduced with the guidance of expression cues. Specifically, with our prior knowledge regularization, the prediction results are more similar to the ground truth than those by the baseline, revealing that our prior knowledge embedding is helpful to capture the correlation between AUs and expressions for AU recognition.

4.4. Qualitative results

To better explain the mechanism of our method, we provide qualitative results based on ResNet50 backbone in this section to answer: **Q3.** Whether the network can learn the logical knowledge between expressions and AUs? **Q4.** Whether the binary relation knowledge of AUs can be learnt by our method? **Q5.** How does prior knowledge help AU recognition?

A3. Visualization of propositional distributions in logic spaces. In order to study the rationality of our learned prior knowledge, we use t-SNE to visualize the distributions of satisfiable and unsatisfiable

Table 1
F1-scores (%) achieved for 12 AUs on BP4D. The best, second best, and third best results of each column are indicated with bold font, brackets, and underline, respectively (see [3,6–10,29–41]).

Method	Publication	AU												Avg.
		1	2	4	6	7	10	12	14	15	17	23	24	
EAC-Net [29]	TPAMI2018	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
JAA-Net [9]	ECCV2018	47.2	44.0	54.9	77.5	74.6	84.0	86.9	61.9	43.6	60.3	42.7	41.9	60.0
LP-Net [30]	CVPR2019	43.4	38.0	54.2	77.1	76.7	83.8	87.2	63.3	45.3	60.5	48.1	54.2	61.0
ARL [31]	TAC2019	45.8	39.8	55.1	75.7	77.2	82.3	86.6	58.8	47.6	62.1	47.4	55.4	61.1
SEV-Net [10]	CVPR2021	**58.2**	50.4	58.3	**81.9**	73.9	**87.8**	87.5	61.6	52.6	62.2	44.6	47.6	63.9
FAUDT [32]	CVPR2021	51.7	49.3	61.0	77.8	79.5	82.9	86.3	[67.6]	51.9	63.0	43.7	[56.3]	64.2
Li et al. [33]	ACMM2021	54.0	46.0	55.7	79.4	78.8	84.5	87.0	67.0	[55.6]	63.1	50.7	55.3	[64.8]
GeoCNN [34]	PR2022	48.4	44.2	59.9	78.4	75.6	83.6	86.7	65.0	53.0	64.7	49.5	54.1	63.6
FAN-Trans [35]	WACV2023	55.4	46.0	59.8	78.7	77.7	82.7	88.6	64.7	51.4	**65.7**	[50.9]	56.0	[64.8]
CL-ILE [36]	BMVC2023	55.1	[52.1]	55.0	78.2	75.5	83.4	88.1	67.4	51.9	59.5	46.9	**62.2**	64.6
Cui et al. [37]	CVPR2023	[57.4]	**52.6**	**64.6**	79.3	**81.5**	82.7	85.6	**67.9**	47.3	58.0	47.0	44.9	64.1
MAL [38]	TAC2023	47.9	49.5	52.1	77.6	77.8	82.8	88.3	66.4	49.7	59.7	45.2	48.5	62.2
AU-FAN [39]	Proc.IEEE2023	54.2	44.9	[61.5]	76.8	76.6	83.6	[88.8]	63.9	52.3	**65.7**	48.5	48.0	63.8
JAO [40]	PRL2024	54.4	50.1	57.7	79.0	76.0	83.7	87.7	64.8	47.9	62.3	44.1	48.4	63.0
SRERL [7]	AAAI2019	46.9	45.3	55.6	77.1	78.4	83.5	87.6	63.9	52.2	63.9	47.1	53.3	62.9
AU-GCN [41]	MMM2020	46.8	38.5	60.1	80.1	79.5	[84.8]	88.0	67.3	52.0	63.2	40.9	52.8	62.8
UGN-B [3]	AAAI2021	54.2	46.4	56.8	76.2	76.7	82.4	86.1	64.7	51.2	63.1	48.5	53.6	63.3
HMP-PS [6]	CVPR2021	53.1	46.1	56.0	76.5	76.9	82.1	86.4	64.8	51.5	63.0	49.9	54.5	63.4
ME-GraphAU [8]	IJCAI2022	53.7	46.9	59.0	78.5	80.0	84.4	87.8	67.3	52.5	63.2	50.6	52.4	64.7
Ours(CNN)	–	54.2	48.1	59.4	[79.8]	[80.3]	84.1	87.7	66.2	54.1	64.4	48.6	55.7	**65.3**
Ours(Transformer)	–	54.4	42.9	60.8	79.4	76.9	83.5	**89.4**	63.2	**55.8**	[65.6]	**52.8**	52.8	[64.8]

Table 2
F1 scores (%) achieved for 8 AUs on DISFA. The best, second best, and third best results of each column are indicated with bold font, brackets, and underline, respectively.

Method	Publication	AU								Avg.
		1	2	4	6	9	12	25	26	
EAC-Net [29]	TPAMI2018	41.5	26.4	66.4	50.7	8.5	**89.3**	88.9	15.6	48.5
JAA-Net [9]	ECCV2018	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
LP-Net [30]	CVPR2019	29.9	24.7	72.7	46.8	49.6	72.9	93.8	65.0	56.9
ARL [31]	TAC2019	43.9	42.1	63.6	41.8	40.0	76.2	[95.2]	[66.8]	58.7
SEV-Net [10]	CVPR2021	55.3	53.1	61.5	53.6	38.2	71.6	**95.7**	41.5	58.8
FAUDT [32]	CVPR2021	46.1	48.6	72.8	[56.7]	50.0	72.1	90.8	55.4	61.5
Li et al. [33]	ACMM2021	47.5	53.3	64.4	51.8	44.4	74.7	92.1	60.7	61.1
GeoCNN [34]	PR2022	**65.5**	**65.8**	67.2	48.6	51.4	72.6	80.9	44.9	62.1
FAN-Trans [35]	WACV2023	56.4	50.2	68.6	49.2	[57.6]	75.6	93.6	58.8	[63.8]
CL-ILE [36]	BMVC2023	58.9	56.4	69.1	**58.5**	54.4	72.2	85.9	47.3	62.8
Cui et al. [37]	CVPR2023	41.5	44.9	60.3	51.5	50.3	70.4	91.3	55.3	58.2
MAL [38]	TAC2023	43.8	39.3	68.9	47.4	48.6	72.7	90.6	52.6	58.0
AU-FAN [39]	Proc.IEEE2023	59.3	55.3	69.4	49.0	45.9	77.0	91.8	60.0	63.5
SRERL [7]	AAAI2019	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
AU-GCN [41]	MMM2020	32.3	19.5	55.7	57.9	**61.4**	62.7	90.9.	60.0	55.0
UGN-B [3]	AAAI2021	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
HMP-PS [6]	CVPR2021	38.0	45.9	65.2	50.9	50.8	76.0	93.3	**67.6**	61.0
ME-GraphAU [8]	IJCAI2022	54.6	47.1	[72.9]	54.0	55.7	76.7	91.1	53.0	63.1
Ours(CNN)	–	53.3	52.1	**77.6**	51.3	48.5	74.6	91.7	54.4	63.0
Ours(Transformer)	–	[62.1]	[61.6]	69.2	47.2	52.6	[78.1]	92.3	53.2	**64.5**

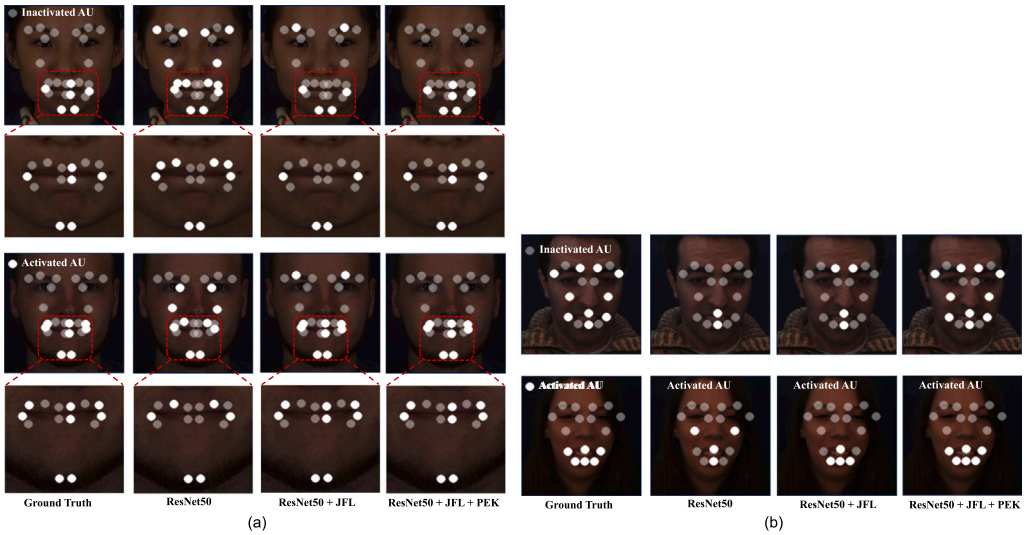

Fig. 5. Visualization of the impact of different modules for two examples from BP4D (a) and DISFA (b). Since some AUs appear at the similar positions, we enlarged the areas near the mouth (the red dotted box) to better visualize the results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

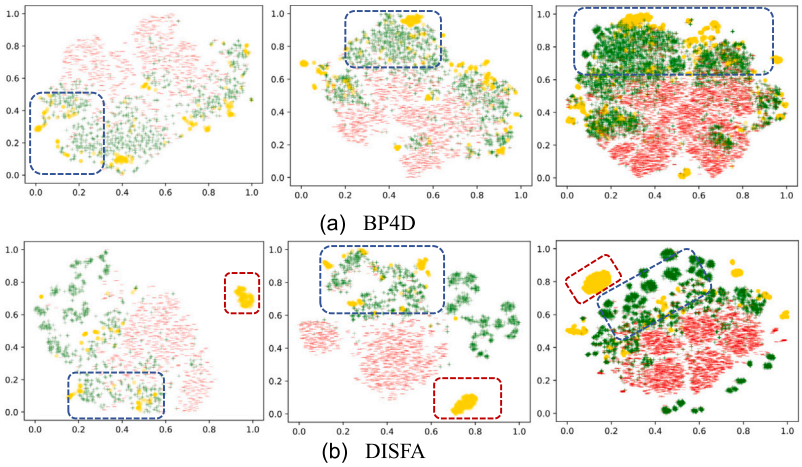

Fig. 6. Distribution visualization of CNF, satisfiable and unsatisfiable propositions in the logic space. Based on the BP4D (a) and DISFA (b), we generate the logic graph (yellow) for each sample, and randomly generate five pairs of satisfiable (green) and unsatisfiable (red) propositions. We use batches of 5, 10, and 50 samples for visualization from left to right. The blue dotted box represents the area where CNF paradigms are close to satisfiable propositions. The red dotted box represents the abnormal area where CNF paradigms gather. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Average AU recognition results of F1-scores (%) achieved by various variants on BP4D and DISFA.

Backbone	BP4D			DISFA		
	Baseline	JFL	JFL+EKR	Baseline	JFL	JFL+EKR
ResNet50	63.3	64.6	**65.4**	61.2	62.1	**63.0**
Transformer	64.1	64.5	**64.8**	63.4	64.2	**64.5**

propositions, as well as CNF logic graphs in Fig. 6, where we use global nodes to visualize proposition distributions. Fig. 6 shows that our logic embedder can push the prior knowledge and the satisfiable propositions closer (blue dotted box in Fig. 6), while well distinguishing the satisfiable and unsatisfiable propositions. This shows that our knowledge embedder has well learned the symbolic knowledge between expressions and AUs, which is helpful for the proposed knowledge regularization.

However, for the DISFA dataset, the logic graphs (yellow) may be clustered together (*e.g.,* red dotted box in Fig. 6). We speculate that this is due to the lack of accurate semantics of the leaf nodes and learnable logic knowledge caused by the sparse labels [42], which may be the reason that our method does not achieve the best result for DISFA based on the structure of the CNN backbone. Whereas, based on the transformer-based backbone, slicing the facial area can better represent the semantic features of AU, thus achieving stronger feature representation capability and help learning logical knowledge. This analysis is validated in Table 2 that our algorithm with the backbone of transformer achieves the best performance on DISFA.

A4. Visualization of the Pearson Correlation Coefficient (PCC) matrices. To study whether our method learns the binary relation knowledge of AUs, the PCC matrices of BP4D and DISFA are visualized in Fig. 7, which are obtained based on the ground-truth AU labels, as well as the binary AU results predicted by the baseline and our method. It shows that our method can better learn the binary relationships between AUs, and outperforms the baseline by the margins of 0.7% and 0.69% in terms of the similarity with the ground truth for BP4D and DISFA, respectively. These results show that our method can learn specific AU relationships for each dataset, besides of the prior knowledge of FACS [2], enabling our method to achieve better generalization performance on real examples.

A5. Visualization of AU activation status before and after knowledge regularization. To discover how the CNF paradigm affects the AU recognition, we visualize the AU activation status before and after knowledge regularization in Fig. 8. It shows that AU–Expression

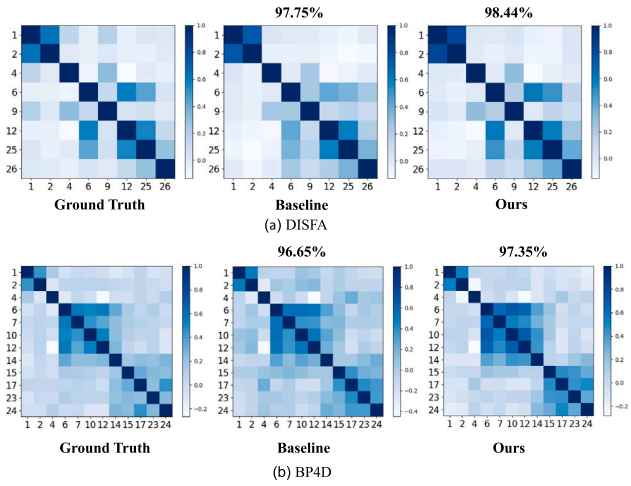

Fig. 7. PCC matrices between AUs for (a): DISFA and (b): BP4D. From left to right, it shows the PCC matrices obtained based on the ground-truth AU labels, by baseline and our method, respectively. Each value above a PCC matrix is the cosine similarity between the vectorized representations of this matrix and the corresponding ground-truth.

knowledge and AU Co-occurrence help recognize unactivated AUs in samples (*e.g.,* 'AU24' in the 1st row, 'AU1' in the 2nd row and 'AU14' in the 3rd row). Meanwhile, the mutually exclusive relationship between AUs helps the network correct mistakenly activated AUs (*e.g.,* 'AU6' in 2nd row). Overall, the explicitly sample-aware prior knowledge in the proposed CNF paradigm can largely reduce false activations and missing activations.

5. Conclusion and discussions

This work proposes a novel AU recognition algorithm called SymGraphAU, to explore the prior and sample-aware logical relationship between AU–AU and AU–Expression that remains underexplored within the existing literature. First, we use joint-task learning to let networks learn facial emotion details at different granularities of AUs and expressions. Subsequently, instead of using unitary prior knowledge for all the samples, we encode AU–Expression relations and tailor a knowledge regularization for each sample via a flexible CNF paradigm. Moreover, We use a regularization mechanism to enable AU recognition to learn

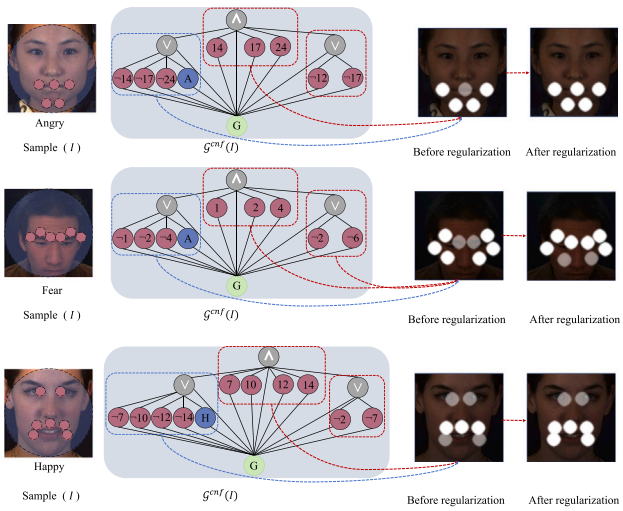

Fig. 8. Prior knowledge visualization for different samples and AU recognition results with our model before and after knowledge regularization. For ease of presentation, we only show the activation of AUs related to the prior knowledge.

the prior knowledge of logic rules from the expressions in terms of graph convolutional networks. Extensive results demonstrate that our method models sample-specific knowledge cues for AU recognition and achieves SOTA F1-score on two mainstream datasets. Ablation studies also validate the effectiveness of each proposed component and demonstrate the competitive recognition performance of our method.

Although our method achieves competitive performances, there is still room for further improvement. First, more accurate semantic nodes should be explored to assist the learning of logical knowledge. Second, for sparsely labeled datasets, *e.g.,* DISFA [27], the prior knowledge generated is also sparse, which leads to insufficient learning of prior knowledge. Therefore, we would explore how to generate richer and more diverse prior knowledge to avoid this insufficient learning. We would further extend our approach and explore its potential applications in a broader range of scenarios.

CRediT authorship contribution statement

Weicheng Xie: Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Fan Yang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Data curation. **Junliang Zhang:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Data curation. **Siyang Song:** Writing – review & editing, Validation, Methodology, Formal analysis. **Linlin Shen:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Zitong Yu:** Writing – review & editing, Methodology, Funding acquisition, Formal analysis. **Cheng Luo:** Writing – review & editing, Visualization, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive suggestions. The work was supported by the National Natural Science Foundation of China under grants no. 62276170, 82261138629, 62306061, the Science and Technology Project of Guangdong Province, China under grants no. 2023A1515011549, 2023A1515010688, the Science and Technology Innovation Commission of Shenzhen, China under grant no. JCYJ20220531101412030, the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), China under Grant No. GML-KF-24-11, Shenzhen Higher Education Stable Support Program General Project, China under Grant 20231120175215001, and Guangdong Provincial Key Laboratory, China under grant no. 2023B1212060076.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patcog.2025.111640.

Data availability

Data will be made available on request.

References

[1] J. Jiang, M. Wang, B. Xiao, J. Hu, W. Deng, Joint recognition of basic and compound facial expressions by mining latent soft labels, Pattern Recognit. 148 (2024) 110173.
[2] P. Ekman, W.V. Friesen, Facial action coding system, Environ. Psychol. Nonverbal Behav. (1978).
[3] T. Song, L. Chen, W. Zheng, Q. Ji, Uncertain graph neural networks for facial action unit detection, in: Proc. AAAI Conf. Artif. Intell., Vol. 35, No. 7, 2021, pp. 5993–6001.
[4] Z. Shao, Y. Zhou, H. Zhu, W.-L. Du, R. Yao, H. Chen, Facial action unit recognition by prior and adaptive attention, Electronics 11 (19) (2022) 3047.
[5] Z. Cui, T. Song, Y. Wang, Q. Ji, Knowledge augmented deep neural networks for joint facial expression and action unit recognition, Proc. Conf. Neural Inf, Process. Syst. 33 (2020) 14338–14349.
[6] T. Song, Z. Cui, W. Zheng, Q. Ji, Hybrid message passing with performance-driven structures for facial action unit detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2021, pp. 6267–6276.
[7] G. Li, X. Zhu, Y. Zeng, Q. Wang, L. Lin, Semantic relationships guided representation learning for facial action unit recognition, in: Proc. AAAI Conf. Artif. Intell., Vol. 33, No. 01, 2019, pp. 8594–8601.
[8] C. Luo, S. Song, W. Xie, L. Shen, H. Gunes, Learning multi-dimensional edge feature-based AU relation graph for facial action unit recognition, in: Int. Joint Conf. Artif. Intell., 2022, pp. 1239–1246.
[9] Z. Shao, Z. Liu, J. Cai, L. Ma, Deep adaptive attention for joint facial action unit detection and face alignment, in: Eur. Conf. Comput. Vis., 2018, pp. 705–720.
[10] H. Yang, L. Yin, Y. Zhou, J. Gu, Exploiting semantic embedding and visual feature for facial action unit detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2021, pp. 10482–10491.
[11] S. Wang, G. Peng, Weakly supervised dual learning for facial action unit recognition, IEEE Trans. Multimed. 21 (12) (2019) 3218–3230.
[12] M. Badea, C. Florea, A. Racoviţeanu, L. Florea, C. Vertan, Timid semi–supervised learning for face expression analysis, Pattern Recognit. 138 (2023) 109417.
[13] Y. Liu, X. Zhang, J. Kauttonen, G. Zhao, Uncertain label correction via auxiliary action unit graphs for facial expression recognition, in: Int. Conf. Pattern Recog., 2022, pp. 777–783.
[14] T. Pu, T. Chen, Y. Xie, H. Wu, L. Lin, Au-expression knowledge constrained representation learning for facial expression recognition, in: Int. Conf. Robot. Autom, 2021, pp. 11154–11161.
[15] S. Eleftheriadis, O. Rudovic, M. Pantic, Multi-conditional latent variable model for joint facial action unit detection, in: IEEE Int. Conf. Comput. Vis., 2015, pp. 3792–3800.
[16] K. Zhao, W.-S. Chu, H. Zhang, Deep region and multi-label learning for facial action unit detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 3391–3399.
[17] Y. Zhang, W. Dong, B.-G. Hu, Q. Ji, Classifier learning with prior probabilities for facial action unit recognition, in: IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 5108–5116.
[18] Y. Xie, Z. Xu, M.S. Kankanhalli, K.S. Meel, H. Soh, Embedding symbolic knowledge into deep networks, Proc. Conf. Neural Inf, Process. Syst. 32 (2019).
[19] D. Yu, B. Yang, D. Liu, H. Wang, S. Pan, A survey on neural-symbolic learning systems, Neural Netw. (2023).
[20] Y. Zhang, X. Chen, Y. Yang, A. Ramamurthy, B. Li, Y. Qi, L. Song, Efficient probabilistic logic reasoning with graph neural networks, in: Int. Conf. Learn. Represent., 2020.

[21] J. Tian, Y. Li, W. Chen, L. Xiao, H. He, Y. Jin, Weakly supervised neural symbolic learning for cognitive tasks, in: Proc. AAAI Conf. Artif. Intell., Vol. 36, No. 5, 2022, pp. 5888–5896.
[22] M. Diligenti, M. Gori, C. Sacca, Semantic-based regularization for learning and inference, Artificial Intelligence 244 (2017) 143–165.
[23] J. Xu, Z. Zhang, T. Friedman, Y. Liang, G. Broeck, A semantic loss function for deep learning with symbolic knowledge, in: Int. Conf. Mach. Learn., PMLR, 2018, pp. 5502–5511.
[24] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: Int. Conf. Learn. Represent., 2017.
[25] A. Ignatiev, A. Morgado, J. Marques-Silva, PySAT: A Python toolkit for prototyping with SAT oracles, in: SAT, 2018, pp. 428–437.
[26] X. Zhang, L. Yin, J.F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, J.M. Girard, Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database, Image Vis. Comput. 32 (10) (2014) 692–706.
[27] S.M. Mavadati, M.H. Mahoor, K. Bartlett, P. Trinh, J.F. Cohn, Disfa: A spontaneous facial action intensity database, IEEE Trans. Affect. Comput. 4 (2) (2013) 151–160.
[28] X. Yin, X. Liu, Multi-task convolutional neural network for pose-invariant face recognition, IEEE Trans. Image Process. 27 (2) (2017) 964–975.
[29] W. Li, F. Abtahi, Z. Zhu, L. Yin, Eac-net: Deep nets with enhancing and cropping for facial action unit detection, IEEE Trans. Pattern Anal. Mach. Intell. 40 (11) (2018) 2583–2596.
[30] X. Niu, H. Han, S. Yang, Y. Huang, S. Shan, Local relationship learning with person-specific shape regularization for facial action unit detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 11917–11926.
[31] Z. Shao, Z. Liu, J. Cai, Y. Wu, L. Ma, Facial action unit detection using attention and relation learning, IEEE Trans. Affect. Comput. 13 (3) (2019) 1274–1289.
[32] G.M. Jacob, B. Stenger, Facial action unit detection with transformers, in: IEEE Conf. Comput. Vis. Pattern Recog., 2021, pp. 7680–7689.
[33] Z. Li, X. Deng, X. Li, L. Yin, Integrating semantic and temporal relationships in facial action unit detection, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 5519–5527.
[34] Y. Chen, G. Song, Z. Shao, J. Cai, T.-J. Cham, J. Zheng, Geoconv: Geodesic guided convolution for facial action unit recognition, Pattern Recognit. 122 (2022) 108355.
[35] J. Yang, J. Shen, Y. Lin, Y. Hristov, M. Pantic, FAN-trans: Online knowledge distillation for facial action unit detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 6019–6027.
[36] T. Lian, D. Adama, P. Machado, D. Vinkemeier, Supervised contrastive learning with identity-label embeddings for facial action unit recognition, in: British Machine Vision Conference, 2023.
[37] Z. Cui, C. Kuang, T. Gao, K. Talamadupula, Q. Ji, Biomechanics-guided facial action unit detection through force modeling, in: IEEE Conf. Comput. Vis. Pattern Recog., 2023, pp. 8694–8703.
[38] Y. Li, S. Shan, Meta auxiliary learning for facial action unit detection, IEEE Trans. Affect. Comput. 14 (3) (2023) 2526–2538.
[39] J. Yang, Y. Hristov, J. Shen, Y. Lin, M. Pantic, Toward robust facial action units' detection, Proc. IEEE 111 (10) (2023) 1198–1214.
[40] Z. Shao, Y. Zhou, F. Li, H. Zhu, B. Liu, Joint facial action unit recognition and self-supervised optical flow estimation, Pattern Recognit. Lett. (2024).
[41] Z. Liu, J. Dong, C. Zhang, L. Wang, J. Dang, Relation modeling with graph convolutional networks for facial action unit detection, in: MultiMedia Modeling: 26th International Conference, 2020, pp. 489–501.
[42] Y. Fan, J. Lam, V. Li, Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution, in: Proc. AAAI Conf. Artif. Intell., 2020, pp. 12701–12708.

Weicheng Xie is currently an associate professor at School of Computer Science and Software Engineering, Shenzhen University, China. He received the B.S. degree in statistics from Central China Normal University in 2008, the M.S. degree in probability and mathematical statistics and Ph.D. degree in computational mathematics from Wuhan University, China in 2010 and 2013. He has been a visiting research fellow with School of Computer Science, University of Nottingham, UK. His current researches focus on facial expression analysis and robust network design.

Fan Yang received the B.Sc. degree from the School of Computer Science and Technology, Tianjin Polytechnic University, in 2021. He is currently pursuing the M.Sc. degree with the Computer Science and Technology, Shenzhen University. His research interests include facial expression recognition and action unit detection.

Junliang Zhang received the B.Sc. degree from the Department of Intelligent Manufacturing of Wuyi University in 2022. He is currently pursuing the M.Sc. degree with the Computer Science and Technology, Shenzhen University. His research interests include facial expression recognition and multi-modal learning in videos.

Siyang Song is currently a Lecturer (Assistant Professor) at the University of Exeter. He is also an affiliated researcher at the Department of Computer Science and Technology, University of Cambridge. His current research interests include affective computing, graph representation learning, computer vision and machine learning.

Linlin Shen is currently a Pengcheng Scholar Distinguished Professor at School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He is also a Honorary professor at School of Computer Science, University of Nottingham, UK. He serves as the Deputy director of National Engineering Lab of Big Data Computing Technology, Director of Computer Vision Institute, AI Research Center for Medical Image Analysis and Diagnosis and China–UK joint research lab for visual information processing. He also serves as the Co-Editor-in-Chief of the IET journal of Cognitive Computation and Systems and Senior Editor of Expert Systems With Applications. His research interests include deep learning, facial recognition, analysis/synthesis and medical image processing. Prof. Shen is listed as the "Most Cited Chinese Researchers" by Elsevier, "Top 0.05% Highly Ranked Scholar" by ScholarGPS, and listed in a ranking of the "Top 2% Scientists in the World" by Stanford University. He received the "Best Paper Runner-up Award" from the journal of IEEE Transactions on Affective Computing, and "Most Cited Paper Award" from the journal of Image and Vision Computing. His cell classification algorithms were the winners of the International Contest on Pattern Recognition Techniques for Indirect Immunofluorescence Images held by ICIP and ICPR. His team has also been the runner-up and second runner-up of a number of competitions for object detection in remote sensing images, nucleus detection in histopathology images and facial expression recognition.

Zitong Yu received the Ph.D. degree in computer science and engineering from the University of Oulu, Oulu, Finland, in 2022. He is currently an Assistant Professor with Great Bay University, China. He was a Postdoctoral Researcher with ROSE Lab, Nanyang Technological University, Singapore. From July to November 2021, he was a Visiting Scholar with TVG, University of Oxford, Oxford, U.K. His research interests include computer vision and biometric security. He was the recipient of the IAPR Best Student Paper Award, IEEE Finland Section Best Student Conference Paper Award 2020, second prize of the IEEE Finland Jt. Chapter SP/CAS Best Paper Award (2022), and World's Top 2% Scientists 2023 by Stanford.

Cheng Luo is currently pursuing his Ph.D. degree at King Abdullah University of Science and Technology, Kingdom of Saudi Arabia. His research interest involves adversarial learning, graph neural network, and video generation. He has published seven CVPR/ICCV/AAAI/IJCAI/ACM MM papers.