Multi-Label Facial Expression Recognition with Graph-Based Modeling of Label Correlation and Node Representation

Miaoxuan Zhang, Weihong Deng, Jing Jiang, Honggang Zhang*

[a]*School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876, China*

Abstract

Multi-label facial expression recognition (ML-FER) is a challenging problem in computer vision and affective computing. Due to the physiological structure of the face and the psychological entanglement of emotions, the basic elements in blended expressions are interrelated. Some expressions are more likely to occur simultaneously, and some rarely appear together. In this work, we propose a Semantic and Visual Multi-Label Relational Graph Convolutional Network (SVML-RGCN) to improve ML-FER by exploiting different relationships between expression labels. Specifically, R-GCN layers are employed to learn both positive and negative relationships between basic expression categories. Additionally, we utilize a combination of semantic and visual information of expressions as the node representation in the graph used for R-GCN. Semantic information is provided by word embeddings, while visual information is derived from basic expression features. We apply Multi-modal Factorized Bilinear Pooling (MFB) to fuse both types of information. Experiments on multiple expression datasets demonstrate the effectiveness of our method. Furthermore, we conduct interpretability analysis of our proposed method through visualization.

Keywords: Relational graph convolutional network (R-GCN), facial expression recognition, multi-label learning, computer vision, affective computing

*Corresponding author. E-mail address: zhhg@bupt.edu.cn (H. Zhang).

1. Introduction

Facial Expression Recognition (FER) is a practical and significant task in the fields of affective computing and computer vision, with applications in areas such as psychotherapy assistance (Candra et al., 2016), human-computer interaction (Gu et al., 2023), and national security (Ben et al., 2023). In the early 1970s, Friesen and Ekman (1978) defined six basic expressions: surprise, fear, disgust, happiness, sadness, and anger. However, expressions in the real world are often more complex and diverse. Previous works (Plutchik, 1991; Ekman, 1984; Hassin et al., 2013; Li and Deng, 2019) have shown that facial expressions are often not pure but a combination of different expression elements, referred to as blended expressions.

Multi-label facial expression recognition (ML-FER) aims to predict a set of basic expression elements that appear simultaneously on a face. This approach is effective in handling complex and blended expressions. However, multi-label classification is inherently more challenging, necessitating the development of more effective classifiers to achieve accurate results.

Due to the physiological structure of the face (*e.g.*, muscle composition) and the psychological entanglement of emotions, there are natural interrelationships between different expression labels. However, previous multi-label expression recognition algorithms (Li and Deng, 2019; Pons and Masip, 2020; Jiang and Deng, 2023) neglected to learn different correlations between labels. Therefore, we attempt to model the correlations between expressions and design a classifier accordingly to improve the performance of multi-label expression recognition.

The graph is an effective structure for portraying correlations. Chen et al. (2019) adapted Graph Convolutional Networks (GCNs) to model relationships between labels, thereby improving the performance of multi-label classification. Nevertheless, the GCN-based multi-label learning approach (ML-GCN) only concerns about the positive relations between labels. However, as for expressions, there are not only positive correlations between distinct expression labels. Some labels seldom appear together, which can be regarded as negative relationships, as demonstrated in Fig. 1. Additionally, ML-GCN denotes node representation with word embeddings of labels. Given the abstract nature of expression labels, relying solely on word embeddings to provide semantic information may be insufficient for effective ML-FER.

In this work, we propose a novel multi-label learning method for ML-FER

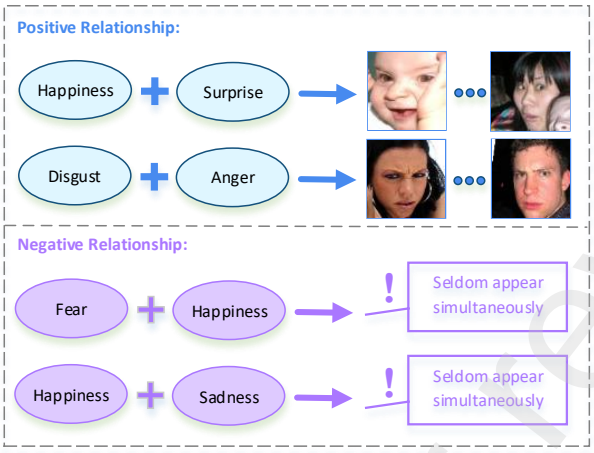

Figure 1: Positive and negative relationships between expression labels.

based on the Relational Graph Convolutional Network (R-GCN) (Schlichtkrull et al., 2018), termed Semantic and Visual Multi-Label R-GCN (SVML-RGCN). Unlike ML-GCN, which only uses GCN (Kipf and Welling, 2016) to model positive relationships between labels, SVML-RGCN considers both positive and negative relations with the help of R-GCN. In that case, a more comprehensive multi-label expression classifier is obtained, as indicated in Fig. 2. Additionally, we use both semantic and visual information as node representations by applying word embeddings and visual features from basic expressions. Multi-modal Factorized Bilinear Pooling (MFB) (Yu et al., 2017) is employed to fuse the visual and semantic information. Our method achieves competitive performance on multiple commonly used blended expression databases. Furthermore, we provide an interpretability analysis of the proposed method.

Our main contributions can be listed as follows:

1. We consider both positive and negative relationships between different expressions and use R-GCN to model these relationships to enhance ML-FER.
2. We combine both visual and semantic information of different expressions using the MFB algorithm to improve node representation in R-GCN.
3. We evaluate our method on the RAF-ML, RAF-Compound, and JAFFE databases, achieving state-of-the-art results. Additionally, we perform

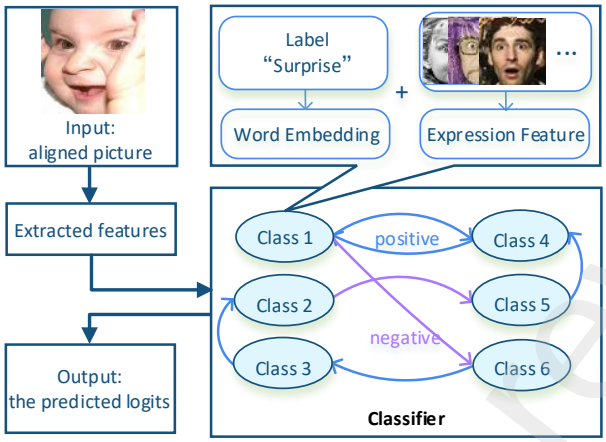

Figure 2: The classifier in proposed SVML-RGCN.

interpretability analysis of the proposed method.

2. Related Work

2.1. Multi-label facial expression recognition

With the enrichment of facial expression datasets and the enhancement of computing ability, FER has been continuously developed. Previous works explored different networks to deal with expression recognition (Zheng et al., 2018; Du and Hu, 2019; Liu et al., 2022; Teng et al., 2023; Zhang et al., 2024; Solis-Arrazola et al., 2024). Additionally, the expression categories have expanded from the limited 6 basic expressions to more complex and diverse expressions. Blended expression refers to a new expression generated by combining more than one basic expression component. Methods to solve blended expression recognition problems can be grouped into two types: methods based on facial action unit (AU) detection and direct multi-label expression recognition approaches.

AU-based approaches are theoretically based on Facial Action Coding System (FACS) (Rosenberg and Ekman, 2005), which divides human faces into a number of independent and interrelated action units (AUs) according to the anatomical characteristics, and analyzed the movement characteristics of these AUs, as well as the expressions related to them. Therefore, researchers are able to analyze various nuanced expressions with the help

of AUs that appear either individually or in combination. Fabian Benitez-Quiroz et al. (2016) proposed a computer vision algorithm to annotate AUs and their intensities from millions of images in the wild. Based on the distributions of AUs, the blended expressions can be identified. Tallec et al. (2022) designed a multi-label detection transformer that utilizes multi-head attention to learn the most relevant parts for each predicted AU in the face image. Kollias (2023) introduced a multi-task learning method C-EXPR-NET for compound expression recognition and AU detection, where the AU detection task is used to enhance facial expression recognition (FER) performance. Unfortunately, there are still very few datasets with AU labels, and the annotation of AUs is also a time-consuming and difficult task. Meanwhile, mapping AU to specific emotional categories is also a challenging task, especially for complex blended expressions.

Directly using multi-label learning methods for blended expression recognition is a more convenient way. The ML-FER methods can intuitively predict the existence of each expression category and quantify its score. Only a few ML-FER models have been developed so far. For instance, Li and Deng (2019) learned the discriminative feature for multi-label expressions by jointly maintaining the local affinity of deep features and the manifold structures of the labels. Li et al. (2021c) proposed a Self-supervised Exclusive-Inclusive Interactive Learning (SEIIL) method for ML-FER in the wild, which includes an emotion disentangling module, an adaptively-weighted ensemble technique, and a conditional adversarial interactive learning module. Pons and Masip (2020) introduced a multi-task and multi-label learning loss function to help FER by sharing a common feature representation with other related tasks. However, these methods didn't concern about different relationships among distinct basic expression elements, in other words, the positive and negative relationships among labels.

In this work, we propose an ML-FER method that uses graph structure to model and explore the label correlation. To be specific, R-GCN (Schlichtkrull et al., 2018) is utilized to capture and learn the relationship among labels, thereafter we use the label correlation information to assist the final classification.

2.2. Graph Convolutional Network and Relational Graph Convolutional Network

Graph Convolutional Network (GCN) (Kipf and Welling, 2016) is modified from convolutional neural networks and has the characteristics of oper-

ating directly on graphs. In GCN, the node representations are updated by propagating information between nodes. Previous works (Yao et al., 2018, 2019) have applied GCN to capture the relationships between different objects and make use of the relationships to improve the performance of models. GCN is also helpful for multi-label classification. Chen et al. (2019) presented a GCN-based multi-label classification model, ML-GCN, which utilized GCN to model the label dependencies and thereafter improve the recognition performance. ML-GCN built the graph over the object labels, and each node representation is constructed by word embeddings. However, in ML-GCN, only the positive relationship between labels is concerned. Additionally, word embeddings can only provide limited weak semantic information, which restricts the performance of ML-GCN.

Since GCN is designed to be used in homogeneous graphs, it does not do well in multi-relational data. In order to utilize both positive and negative correlation between labels, Relational Graph Convolutional Network (R-GCN) (Schlichtkrull et al., 2018) is worth considering. R-GCN solves the problem of using GCN to handle the impact of different edge relationships on nodes in graph structures. Xu and Yang (2019) proposed an end-to-end coreference resolver by combining pre-trained Bidirectional Encoder Representations from Transformers (BERT) with R-GCN, where R-GCN was used to learn structural syntactic information. Li et al. (2021a) proposed an LSTM Relational Graph Convolutional Network (LSTM-RGCN) model, which used R-GCN to model the positive and negative correlation among stocks. Feng et al. (2021) introduced Bot detection with Relational Graph Convolutional Networks (BotRGCN), which constructed a heterogeneous graph from follow relationships and used R-GCN to learn from it. However, as far as we know, no work has been conducted on modeling expression correlations using R-GCN.

In this work, we adapt ML-GCN to ML-FER task and improve it to suit the task. To be specific, we utilize R-GCN instead of GCN to model both positive and negative relationships between expression labels. In addition, we make use of both semantic information from word embeddings and visual information from basic expression image features when constructing the graph.

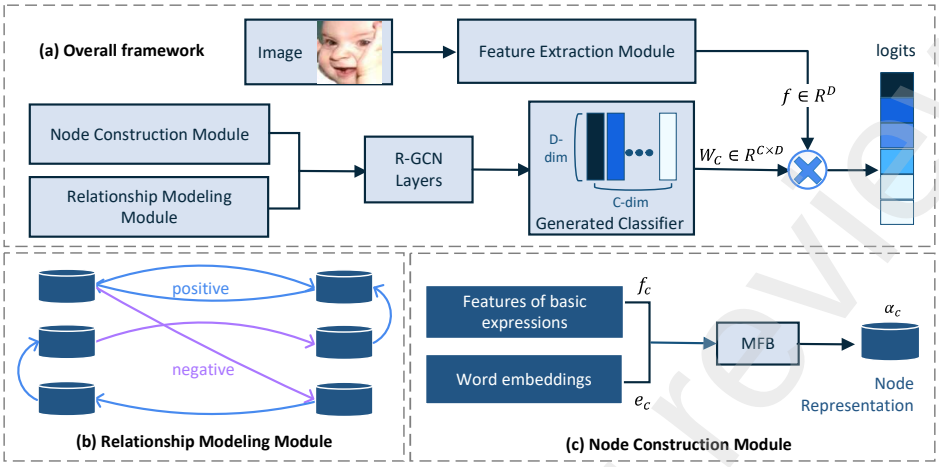

Figure 3: The proposed model SVML-RGCN. (a) is the overall framework of the proposed model. Initially, a feature extraction module is employed to extract features from blended expression images. Then a classifier is learnt with the help of R-GCN. (b) and (c) are two main components to construct the graph for R-GCN. (b) is the relationship modeling module, in which the positive and negative correlations of the expression labels are leveraged to build the graph. (c) is the node construction module, where the node representation of the graph comprises visual features and semantic word embeddings of expressions.

3. Approach

In this part, we focus on our SVML-RGCN model for ML-FER. Firstly, we introduce the overall framework of the model. Then we give a detailed description of important module in SVML-RGCN.

3.1. Overall Model

As illustrated in Fig. 3, we propose SVML-RGCN for multi-label facial expression recognition. First, we utilize a feature extraction module to obtain features from input images. This module can be implemented by a generic feature extraction backbone. For simplicity, we use the convolutional modules in ResNet50 network (He et al., 2016). The main contribute of the SVML-RGCN model is to learn an effective multi-label classifier by leveraging R-GCN. Futhermore, the construction of the graph in R-GCN involves two essential parts: edges and nodes. Edges model the relationships between different basic expressions, and nodes represent these expression elements.

To be specific, we construct the edges of the graph by utilizing both positive and negative relationships among the six basic expression elements.

Both types of relationships are modeled in the relationship modeling module. Then, the node representation of the graph is addressed in the node construction module. We extract features from basic expression images to obtain visual information and utilize word embeddings of labels to capture semantic information. These two types of information are then fused using a bilinear pooling method called Multi-modal Factorized Bilinear Pooling (MFB) (Yu et al., 2017). The output of the MFB serves as the node representation of the graph.

After that, R-GCN layers are learned over the label graph and get a classifier $W_C \in R^{C \times D}$, where $C = 6$ is the number of categories and $D = 2048$ is the output dimension of each node in the graph. As shown in Eq. 1, we then multiply W_C by feature $f \in R^D$, which is the output of feature extraction module, to get the logits for classification.

$$\hat{y} = W_C f. \tag{1}$$

The loss function we adopt during the training stage is the traditional multi-label classification loss as follows:

$$\mathcal{L} = \sum_{c=1}^{C} y^c log(\sigma(\hat{y}^c)) + (1 - y^c) log(1 - \sigma(\hat{y}^c)), \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function; $y \in R^C$ is the ground truth label of an image, and $y^c = \{0, 1\}$ indicates whether label L_c appears in the image; $\hat{y} \in R^C$ is the predicted logits of an image.

3.2. Relationship Modeling Module

Due to the physiological structure of the face and the psychological motivations behind emotions, basic expression elements are not completely independent. As demonstrated in Fig. 1, some expressions, such as disgust and anger, have a higher tendency to occur together. This frequent co-occurrence suggests a positive relationship between these expressions, indicating that the activation of one is likely to enhance or coincide with the activation of the other. Conversely, the co-occurrence of some other emotional expressions is much less common. For instance, fear and happiness rarely appear simultaneously, reflecting a negative relationship. This inverse relationship suggests that the presence of one of these emotions is likely to inhibit or diminish the likelihood of the other occurring.

In SVML-RGCN, we utilize both positive and negative relationships between different expressions to construct a helpful classifier. Specifically, we attempt to use the concurring probability of labels (Chen et al., 2019) to model those relationships. Suppose the six basic expression categories are denoted as $\{c_1, c_2, ..., c_6\}$. We set M_{ij} to count the concurring times of label L_{c_i} and L_{c_j}, in other words, the concurring times of basic expressions c_i and c_j. Then the probability P_{ij} is set to be

$$P_{ij} = \frac{M_{ij}}{N_{c_i}}, \tag{3}$$

where N_{c_i} denotes the number of samples with label L_{c_i}.

We model the relationships $\mathcal{R}$ between labels according to probability P_{ij}. As shown in Eq. 4, the positive relationship represents that two labels have a higher probability to appear simultaneously, while the negative relationship means two labels have a lower probability to appear concurrently. t_1 and t_2 are the thresholds used to define the correlation relationships of labels. Expression labels with a co-occurrence probability lower than t_1 are considered to have a negative correlation, while those with a co-occurrence probability higher than t_2 are viewed as positively correlated. Co-occurrence probabilities between t_1 and t_2 indicate no significant correlation relationship between the labels.

$$\mathcal{R} = \begin{cases} negative, & \text{if } P_{ij} < t_1 \\ positive, & \text{if } P_{ij} > t_2 \end{cases}. \tag{4}$$

3.3. Node Construction Module

To obtain more meaningful node representations and thereby enhance the encoding of diverse basic expressions, we leverage two types of information: semantic and visual information. These are subsequently fused using the MFB algorithm.

Semantic Information. Semantic information is derived from word embedding vectors, which convert words into numerical vectors that encapsulate their semantic relationships. In SVML-RGCN, we employ word embedding vectors trained by GloVe (Pennington et al., 2014) on Wikipedia. The word vectors of the six basic expression labels (Friesen and Ekman, 1978), which are surprise, fear, disgust, happiness, sadness, and anger, are selected to provide semantic information.

Visual Information. Since facial expressions are inherently abstract, relying solely on the semantic information provided by word embeddings for node representations may be insufficient. Therefore, we incorporate visual information alongside semantic information to represent each facial expression label as a node. As deep neural network can extract visual features of facial expressions from basic expression images, we utilize the network to obtain and provide visual information of different expressions. Let C denote the number of categories. The extracted feature for each category can be represented as

$$f_c = \frac{\sum_{n=1}^{N_c} \varphi(I_{c,n})}{N_c}, \tag{5}$$

where φ is the feature extractor; In SVML-RGCN, the ResNet50 model trained on basic expression data is used to extract basic expression features from images. $I_{c,n}$ represents one of the basic-expression image with the c^{th} category ($1 \leq c \leq C$). N_c is the number of images with category c, which are used for feature extraction.

Feature Fusion. MFB is a variant of bilinear pooling (Lin et al., 2015) technology, which can be used to fuse different features and output vectors. In the proposed method, we adopt MFB to combine visual and semantic information. Assuming $\{f_c\}_{c=1}^{C}$ are the basic-expression extracted features, and $\{e_c\}_{c=1}^{C}$ are the word embeddings of labels, the main process can be shown in Eq.6:

$$\alpha_c = \text{SumPooling}(W^T f_c \circ V^T e_c, k), \tag{6}$$

where $\text{SumPooling}(x, k)$ would perform the sum pooling process with a one-dimensional window of size k. $\circ$ is the element-wise multiplication of two vectors. $W \in R^{d_1 \times ko}$ and $V \in R^{d_2 \times ko}$ are parameter matrices that need to be learned. d_1 is the dimension of f_c, d_2 is the dimension of e_c, and o is the dimension of output vector α_c.

$\{\alpha_c\}_{c=1}^{C}$, containing semantic and visual information, will be provided as the node representation of R-GCN.

3.4. R-GCN Layers

After constructing the edges and nodes of the graph, we apply R-GCN layers, which are designed to deal with multi-relational data, to learn a classifier. The propagation rule of R-GCN is

$$h_i^{(l+1)} = \sigma\Big(\sum_{r \in \mathcal{R}} \sum_{j \in N_i^r} \frac{1}{c_{i,r}} W_r^{(l)} h_j^{(l)} + W_0^{(l)} h_i^{(l)}\Big), \tag{7}$$

in which $h_i^{(l)} \in R^{d^{(l)}}$ is the feature description of node i ($1 \leq i \leq C$) in layer l, $d^{(l)}$ is the dimensionality of node feature in this layer. For the input layer, $h_c^{(0)} = \alpha_c$ ($1 \leq c \leq C$). $\mathcal{R}$ denotes the relations defined in Section 3.2. N_i^r is the set of neighbor indices of node i in relation $r \in \mathcal{R}$. $c_{i,r}$ is a normalization constant. In SVML-RGCN, $c_{i,r} = |N_i^r|$. $W_r \in R^{d^{(l+1)} \times d^{(l)}}$ and $W_0 \in R^{d^{(l+1)} \times d^{(l)}}$ are matrix that need to be learn. $\sigma(\cdot)$ is the activation function.

Based on the R-GCN network, we exploit the semantic and visual information of expression labels and model the positive and negative correlations between different expressions to learn an effective multi-label expression classifier.

4. Experiments

4.1. Evaluation Metrics

Following (Li and Deng, 2019; Jiang and Deng, 2023), we utilize 9 metrics to evaluate our proposed method, including Hamming Loss (HL), Coverage, One-Error (OE), Ranking Loss (RL), Average Precision (AP), Micro-F1, Macro-F1, Micro-AUC and Macro-AUC. For the former four metrics, the smaller is the better; while for the other metrics, the larger is the better (It is indicated by "↓" and "↑" respectively in all tables and figures). For each image, the labels will be predicted as true if the confidence scores are greater than 0.5.

Given that C is the number of labels, N is the number of samples, $y_{c,n}$ is the ground truth label of the n^{th} image with the c^{th} label, and $\hat{y}_{c,n}$ is the prediction. We will briefly describe the nine evaluation metrics as follows.

Hamming Loss (HL) is the fraction of incorrect labels to the total number of labels:

$$\frac{1}{NC} \sum_{n=1}^{N} \sum_{c=1}^{C} \mathbb{I}(y_{c,n} \neq \hat{y}_{c,n}), \tag{8}$$

where $\mathbb{I}$ is the indicator function.

Coverage can examine the depth of search required to cover all the labels belonging to the sample in the ranking queue of category scores:

$$\frac{1}{N} \sum_{n=1}^{N} \max_{c:y_{c,n}=1} |\{k : \hat{y}_{k,n} \geq \hat{y}_{c,n}\}|, \tag{9}$$

where $|\cdot|$ is the cardinality of a set.

One-Error (OE) describes the extent to which the top-ranked predicted label is not in the set of groud-truth labels.

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{I}(\hat{y}_{0,n} \notin \{c : y_{c,n} = 1\}), \tag{10}$$

where $\hat{y}_{0,n}$ is the top-ranked label in predicted labels $\hat{y}_n$ of sample n.

Ranking Loss (RL) examines the extent to which irrelevant categories are located before relevant labels in the ranking queue.

$$\frac{1}{N}\sum_{n=1}^{N}\frac{|\{(c,k) : \hat{y}_{c,n} < \hat{y}_{k,n}, y_{c,n} = 1, y_{k,n} = 0\}|}{|Y_n|(C - |Y_n|)}, \tag{11}$$

where $|Y_n|$ denotes the number of true labels that sample n contains.

Average Precision (AP) reflects the mean fraction of relevant labels that are ranked above another relevant label.

$$\frac{1}{N}\sum_{n=1}^{N}\frac{1}{|Y_n|}\sum_{c:y_{c,n}=1}\frac{|\{k : y_{c,n} = 1, \hat{y}_{k,n} \geq \hat{y}_{c,n}\}|}{|\{k : \hat{y}_{k,n} \geq \hat{y}_{c,n}\}|}. \tag{12}$$

Micro-F1 represents the micro-averaged F1 score. The F1 score is the weighted harmonic mean value of precision and recall. Micro-averaging indicates calculating the global metrics (True Positive, False Positive and False Negative) for all categories.

Macro-F1 is the macro-averaged F1 score, implying that it first calculates the F1 score for each category separately, then takes its average.

Micro-AUC represents the micro-calculated AUC. AUC denotes the area under the receiver operating characteristic (ROC) curve. Micro-AUC accumulates all confusion matrices together and then calculates the AUC.

Macro-AUC denotes the macro-calculated AUC. Macro-AUC calculates the metrics for each class individually and then takes the arithmetic average of these values to obtain the final result.

4.2. Experimental Conditions

Datasets. We conducted experiments on three commonly used blended facial expression datasets, including RAF-ML, RAF-Compound and JAFFE dataset. Examples of images from these datasets are depicted in Fig. 4.

RAF-ML (Li and Deng, 2019) is a multi-label expression database that contains 4908 real-world images with blended emotions. The database is annotated with six-dimensional basic expression labels by 315 well-trained annotators. During the experiment, the database is split into a training set with 3926 images and a test set with 982 images.

RAF-DB (Li et al., 2017) is a well-known large facial expression database with around 30,000 real-world facial images downloaded from the Internet. In the basic subset of RAF-DB (RAF-Basic), there are 7 labels, including 6 basic expressions and neutral. In our experiment, only the images labeled with 6 basic expressions in the training set (9747 images in total) are applied. We extracted basic expression features from the images in RAF-Basic for the proposed model.

RAF-Compound is a compound subset of RAF-DB, which includes 3954 images with 11 kinds of compound expressions, such as happily-surprised, sadly-fearful. Since we use multi-label learning in our proposed method, we convert the original labels in RAF-compound into multi-label format. To be specific, each compound expression label is converted into 6-dimensional basic expression labels $y \in R^C$. If the original label covers category c, $y^c = 1$; otherwise $y^c = 0$. Additionally, in the experiment, 3162 images are divided into a training set and 792 images are divided into a test set.

JAFFE (Lyons et al., 1998) is a lab-controlled database with 213 images from 10 Japanese females. The dataset provides emotion intensity distribution for each sample. In this experiment, we use a threshold of 3 to convert the expression intensity distribution into multiple labels. Furthermore, considering the small number of images in JAFFE, we applied 5-fold cross-validation to evaluate the performance of the model.

Implementation Details. All the images are aligned according to two eye centers and the mouth center, then resized to 224 $\times$ 224, as shown in Fig. 4. Random horizontally flipping of images is applied as data augmentation. We adapt the ResNet50 network trained on RAF-Basic to extracted features from basic expression images. Each output feature has a dimensionality of 2048. The convolutional layers of the pretrained ResNet50 model are also used as the feature extraction module of the proposed model. Unless otherwise stated, we adapt 2 R-GCN layers, both of which have an output dimensionality of 2048. We train the model using a GeForce GTX 1080 Ti GPU, with batch size of 32. The Adam optimizer is used and the weight decay is set to 2×10^{-4}. The network is trained for 80 epochs in total, and the learning rate decays by a factor of 10 for every 20 epochs. Due to variations

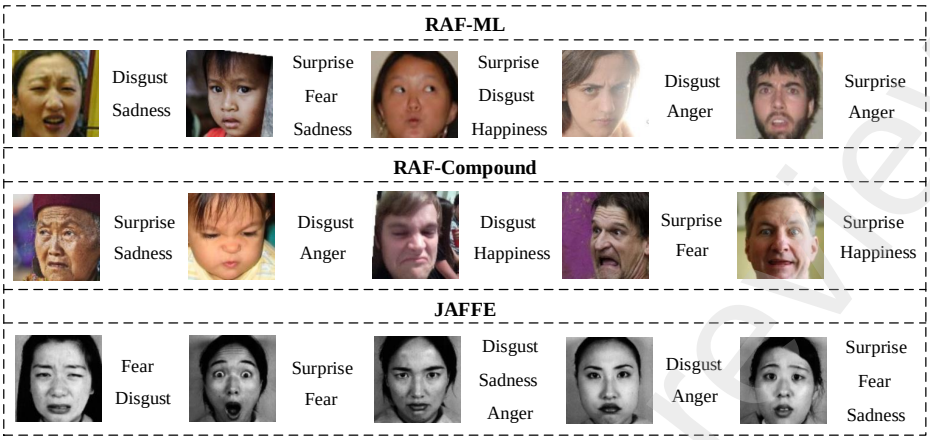

Figure 4: Some example images in RAF-ML (top), RAF-Compound (middle) and JAFFE (bottom) datasets. We transformed the labels of images in the datasets into a multi-label format.

in collection environments and data distributions across datasets, some hyperparameters are adjusted accordingly. In RAF-ML and RAF-Compound, the initial learning rate of the classifier branch is set to 2×10^{-4}, while it is 2×10^{-6} in feature extraction backbone. The negative and positive thresholds t_1, t_2 are 0.15 and 0.4, respectively. Regarding k and o associated with the MFB process, in RAF-ML they are set to 4 and 1024, respectively, while in RAF-Compound they are set to 5 and 2000, respectively. Additionally, in JAFFE, the initial learning rates for the classifier and feature extraction branches are 5×10^{-4} and 1×10^{-4}, respectively. t_1 is set to 1×10^{-6}, a decimal that is close to but greater than zero. t_2 equals 0.45. k and o are set to 5 and 1000, respectively.

4.3. Experimental Results

4.3.1. Comparisons with State-of-the-Art Methods

We report the quantitative results of our proposed results and state-of-the-art (SOTA) methods on datasets RAF-ML, RAF-Compound and JAFFE. Firstly, we compare SVML-RGCN on RAF-ML with other methods, such as DBM-CNN (Li and Deng, 2019), ViT (Dosovitskiy et al., 2020), SEIIL (Li et al., 2021c), MF-JLE (Li et al., 2021b), CMCNN (Yu and Xu, 2022), and D^2S (Wang et al., 2022), as shown in Table 1(a). Notice that the methods with '+MLkNN' indicate they use MLkNN algorithm to classify the learned

expression representation separately. From the results, it is obvious that our method outperforms SOTA methods, which proves the effectiveness of the proposed method. Additionally, we also compare SVML-RGCN with other methods on RAF-Compound in Table 1(b), including ViT (Dosovitskiy et al., 2020), SEIIL (Li et al., 2021c), CMCNN (Yu and Xu, 2022), and VAC (Guo et al., 2019). The results indicate that our method is better than the existing methods in most metrics. We similarly perform comparisons on dataset JAFFE. The results on JAFFE are shown in Table 1(c). We compare with methods that use multi-label learning on JAFFE, including DBM-CNN (Li and Deng, 2019) and DBM-DACNN (Li and Deng, 2019), VAC (Guo et al., 2019), EAC (Zhang et al., 2022), and AFC+SAD (Jiang and Deng, 2023). Notice that DBM-DACNN is based on DBM-CNN and employs domain adaption from dataset RAF-ML to JAFFE. Our method surpasses all these methods, demonstrating the effectiveness of our approach in different datasets. Since blended expressions are inherently complex and subtle, and the amount of existing data is still limited, improving the performance of multi-label facial expression recognition tasks is relatively challenging. However, our method still outperforms SOTA methods, improving its effectiveness.

4.3.2. Ablation Studies

Comparison with the baseline. We regard the ResNet50 model (He et al., 2016) as baseline since we use the convolutional layers of ResNet50 in feature extraction module. Compared with our proposed SVML-RGCN, ResNet50 can be seen as using an automatic learning classifier after the convolutional layers. On the contrary, our method applies a carefully designed classifier that not only has better interpretability for humans, but also improves recognition performance. We compare our proposed method SVML-RGCN with the baseline in Table 2. The results demonstrate that compared with the baseline (for a fair comparison, all the settings are the same), our model performs better on all evaluation metrics, which proves the effectiveness of SVML-RGCN.

Is positive relationship enough? In this part, we explore whether it is sufficient to employ only positive relationships between labels. In other words, is it necessary to model both positive and negative correlations? We compare the model that uses only positive relations (named SVML-GCN) and the proposed SVML-RGCN, which uses not only positive but also neg-

Table 1: Comparisons with state-of-the-arts on RAF-ML, RAF-Compound and JAFFE datasets. Bold indicates the best result under each evaluation index.

(a) Results on RAF-ML

Methods	HL↓	Cov↓	OE↓	RL↓	AP↑	Micro-F1↑	Macro-F1↑	Micro-AUC↑	Macro-AUC↑
DBM-CNN+MLkNN (Li and Deng, 2019)	0.168	1.965	0.133	0.128	0.873	0.768	0.739	0.901	0.891
ViT+MLkNN (Dosovitskiy et al., 2020)	0.170	2.076	0.127	0.141	0.865	0.764	0.742	0.878	0.864
SEIIL (Li et al., 2021c)	0.156	**1.830**	0.094	0.100	0.896	0.785	0.764	0.830	0.818
SEIIL+MLkNN (Li et al., 2021c)	0.161	1.856	0.108	0.106	0.890	0.786	0.767	0.912	0.908
MF-JLE (Li et al., 2021b)	0.155	1.879	0.123	0.111	0.828	0.783	0.762	0.909	0.897
CMCNN (Yu and Xu, 2022)	0.180	2.006	0.134	0.129	0.870	0.751	0.728	0.891	0.880
D^2S (Wang et al., 2022)	0.171	1.971	0.117	0.124	0.882	0.770	0.752	0.897	0.887
Ours	**0.146**	**1.830**	**0.090**	**0.099**	**0.901**	**0.798**	**0.778**	**0.919**	**0.911**

(b) Results on RAF-Compound

Methods	HL↓	Cov↓	OE↓	RL↓	AP↑	Micro-F1↑	Macro-F1↑	Micro-AUC↑	Macro-AUC↑
ViT+MLkNN (Dosovitskiy et al., 2020)	0.149	1.684	0.128	0.107	0.877	0.768	0.738	0.823	0.803
SEIIL (Li et al., 2021c)	0.129	1.641	0.097	0.098	0.893	**0.802**	0.778	0.849	0.831
SEIIL+MLkNN (Li et al., 2021c)	0.139	1.698	0.114	0.106	0.884	0.798	0.777	0.922	0.912
CMCNN (Yu and Xu, 2022)	0.152	1.788	0.151	0.127	0.862	0.765	0.743	0.904	0.895
VAC (Guo et al., 2019)	**0.128**	1.645	0.097	0.097	0.893	0.800	0.772	0.927	0.920
Ours	0.133	**1.606**	**0.095**	**0.093**	**0.894**	0.798	**0.779**	**0.931**	**0.925**

(c) Results on JAFFE

Methods	HL↓	Cov↓	OE↓	RL↓	AP↑	Micro-F1↑	Macro-F1↑	Micro-AUC↑	Macro-AUC↑
DBM-CNN+MLkNN (Li and Deng, 2019)	0.169	1.415	0.138	0.089	0.891	–	–	–	–
DBM-DACNN+MLkNN (Li and Deng, 2019)	0.161	1.369	0.127	0.082	0.906	–	–	–	–
VAC (Guo et al., 2019)	0.181	1.608	0.186	0.120	0.863	0.688	0.617	0.891	0.908
EAC (Zhang et al., 2022)	0.166	1.452	0.130	0.088	0.897	0.709	0.669	0.903	0.910
AFC+SAD (Jiang and Deng, 2023)	0.153	1.392	**0.099**	0.081	0.912	0.748	0.723	0.910	0.903
Ours	**0.129**	**1.335**	0.113	**0.067**	**0.916**	**0.798**	**0.780**	**0.932**	**0.930**

ative relations, as shown in Table 2. To be specific, since GCN is sufficient to model the positive relations between labels, in SVML-GCN, we use GCN layers to learn positive relations, just as ML-GCN (Chen et al., 2019) does. The only difference from the original ML-GCN is that the input node representations of the first GCN layer are still a combination of semantic and visual information, and this is for a fair comparison with SVML-RGCN. Following (Chen et al., 2019), we apply the correlation matrix for GCN and use the re-weighted scheme in SVML-GCN, as shown in Eq. 13 and Eq. 14. Note that A is the binary correlation matrix and $\widetilde{A}$ is the re-weighted correlation matrix. The threshold t used in SVML-GCN is equal to the positive threshold t_2 in SVML-RGCN. And p in Eq.14 is set to 0.2, the same as in ML-GCN.

$$A_{ij} = \begin{cases} 0, & \text{if } P_{ij} < t \\ 1, & \text{if } P_{ij} > t \end{cases}, \tag{13}$$

Table 2: Comparisons with the baseline, SVML-GCN and SVML-RGCN. Bold indicates the best result under each evaluation index.

Datasets	Methods	Positive relations	Negative relations	HL↓	Cov↓	OE↓	RL↓	AP↑	Micro-F1↑	Macro-F1↑	Micro-AUC↑	Macro-AUC↑
RAF-ML	Baseline	×	×	0.1511	1.8371	0.0937	0.1012	0.8989	0.7872	0.7689	0.9173	0.9099
	SVML-GCN	✓	×	0.1514	1.8320	0.0937	0.1008	0.8984	0.7880	0.7707	0.9173	0.9104
	SVML-RGCN	✓	✓	**0.1456**	**1.8299**	**0.0896**	**0.0988**	**0.9011**	**0.7975**	**0.7779**	**0.9189**	**0.9108**
RAF-Compound	Baseline	×	×	0.1362	1.6326	**0.0947**	0.0964	0.8906	0.7900	0.7679	0.9309	0.9242
	SVML-GCN	✓	×	0.1351	1.6225	0.0972	0.0952	0.8913	0.7928	0.7719	0.9305	0.9240
	SVML-RGCN	✓	✓	**0.1326**	**1.6060**	**0.0947**	**0.0931**	**0.8939**	**0.7982**	**0.7788**	**0.9312**	**0.9249**
JAFFE	Baseline	×	×	0.1464	1.4160	0.1563	0.0837	0.8932	0.7691	0.7468	0.9220	0.9225
	SVML-GCN	✓	×	0.1389	1.4070	0.1286	0.0782	0.9022	0.7726	0.7477	0.9249	0.9143
	SVML-RGCN	✓	✓	**0.1291**	**1.3346**	**0.1128**	**0.0671**	**0.9163**	**0.7982**	**0.7797**	**0.9316**	**0.9300**

Table 3: Ablation study of semantic and visual information in SVML-RGCN. Bold indicates the best result under each evaluation index.

Datasets	Methods	Semantic information	Visual information	HL↓	Cov↓	OE↓	RL↓	AP↑	Micro-F1↑	Macro-F1↑	Micro-AUC↑	Macro-AUC↑
RAF-ML	SML-RGCN	✓	×	0.1511	1.8422	**0.0896**	0.1012	0.8990	0.7885	0.7698	0.9176	0.9097
	VML-RGCN	×	✓	0.1488	1.8340	0.0916	0.1004	0.8990	0.7920	0.7732	0.9180	0.9103
	SVML-RGCN	✓	✓	**0.1456**	**1.8299**	**0.0896**	**0.0988**	**0.9011**	**0.7975**	**0.7779**	**0.9189**	**0.9108**
RAF-Compound	SML-RGCN	✓	×	0.1359	1.6073	**0.0985**	**0.0931**	0.8935	0.7917	0.7710	0.9314	0.9246
	VML-RGCN	×	✓	**0.1326**	1.616	**0.0947**	0.0938	0.893	0.7981	0.7782	**0.9316**	0.9247
	SVML-RGCN	✓	✓	**0.1326**	**1.6060**	0.0947	**0.0931**	**0.8939**	**0.7982**	**0.7788**	0.9312	**0.9249**
JAFFE	SML-RGCN	✓	×	0.1368	1.4605	0.1227	0.0889	0.8999	0.7790	0.7571	0.9163	0.9181
	VML-RGCN	×	✓	0.1455	1.3426	0.1448	0.0766	0.9024	0.7644	0.7522	0.9240	0.9230
	SVML-RGCN	✓	✓	**0.1291**	**1.3346**	**0.1128**	**0.0671**	**0.9163**	**0.7982**	**0.7797**	**0.9316**	**0.9300**

$$\widetilde{A} = \begin{cases} (p/\sum_{j=1,i\neq j}^{C} A_{ij}) \times A_{ij}, & \text{if } i \neq j \\ 1-p, & \text{if } i = j \end{cases}. \tag{14}$$

As illustrated in Table 2, compared to SVML-GCN, SVML-RGCN reports better performance on all evaluation indices, indicating the validity and necessity of applying both positive and negative relations between labels.

Effect of semantic and visual information. In this part, we separately examine the effects of semantic and visual information on our model. Specifically, we explore the model with only word embeddings as node features (named SML-RGCN) and the model with only basic expression features (named VML-RGCN). The former only utilizes semantic information, while the latter only uses visual information. We compare SML-RGCN, VML-RGCN, and SVML-RGCN in three databases, and the results are shown in Table 3. As shown, although the performances of SML-RGCN and VML-RGCN are better than the baseline in most cases, SVML-RGCN still outperforms both on most metrics. This illustrates the necessity for models to use not only semantic but also visual information. Additionally, we note

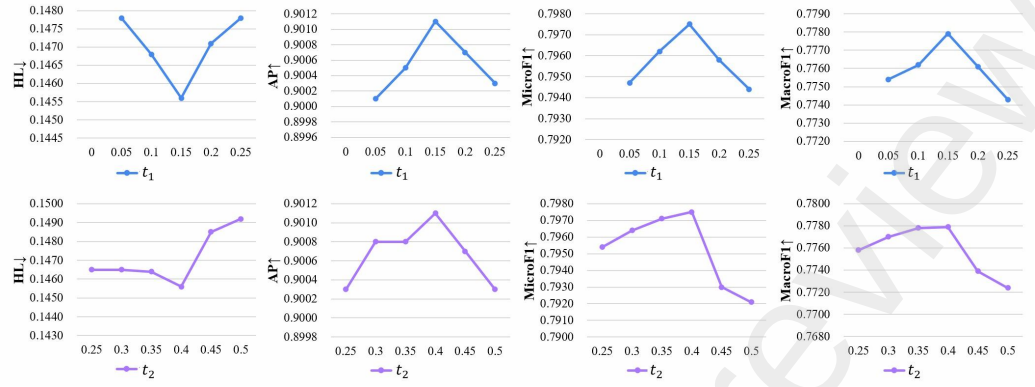

Figure 5: Performance comparisons with different relation thresholds t_1 (top), t_2 (bottom).

that the performance of VML-RGCN is slightly better than that of SML-RGCN, indicating that visual information contributes more to classification than semantic information.

Effect of different thresholds t_1, t_2.

To evaluate the effect of different thresholds, we change negative threshold t_1 in a set of $\{10^{-6}, 0.05, 0.1, 0.1, 0.2, 0.25\}$, and positive threshold t_2 in a set of $\{0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$. First we observe the outcomes in unconstrained real-world expression datasets. Take the RAF-ML dataset as an example, the results are shown in Fig. 5. Due to the absence of two labels with a concurring probability of 0 in RAF-ML training set, edges with negative relationships do not exist when $t_1 = 10^{-6}$. Therefore, $t_1 = 10^{-6}$ is invalid for RAF-ML. According to the results, the optimal performance on RAF-ML is obtained when $t_1 = 0.15$ and $t_2 = 0.4$. A similar trend curve to Fig. 5 can also be observed on the RAF-Compound dataset, with the best performance achieved at the same thresholds. To provide a more comprehensive perspective, we also conduct experiments on a laboratory-controlled dataset, JAFFE. On JAFFE, the model achieves the best performance when $t_1 = 10^{-6}$ and $t_2 = 0.45$. The difference in thresholds may be caused by significant differences in the data volume and sampling environment of the datasets. This also demonstrates that the distribution of expression data is noticeably different between the laboratory and real-world environments.

Effect of R-GCN layers

To explore the model's performance with different numbers of R-GCN layers, we conduct experiments on models with 1, 2 and 3 R-GCN layers. The output dimensionality of all layers is set to 2048. The results are listed

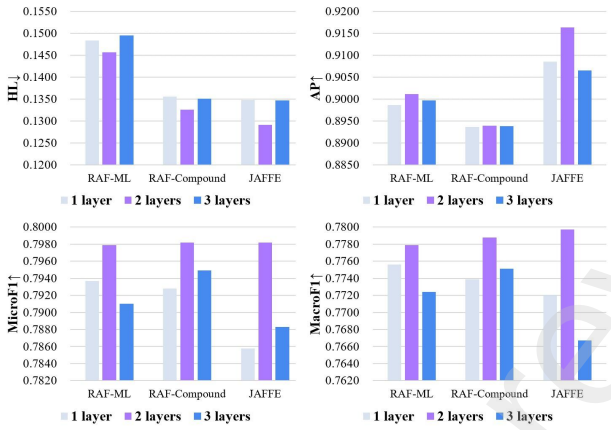

Figure 6: Comparisons with different depths of R-GCN in our model.

in Fig. 6. As shown, when there are two R-GCN layers, the model gets optimal performance. If there is only one R-GCN layer, the model may not fully learn the relationships between different labels. However, when using three R-GCN layers, there may be an issue of over-smoothing, which affects the model's performance.

4.4. Interpretability Analysis

Compared to the baseline, ResNet50 model with an automatic learning classifier, our approach not only has more competitive performance but is also more interpretable. This is because our model uses a well-designed classifier based on semantic and visual features of different expressions, as well as correlations between them. Therefore, it is more interpretable than classifiers that learn parameters automatically, since it is hard for human to understand the basis on which the automatically learned classifier makes its classifications.

Additionally, we used Grad-CAM (Selvaraju et al., 2017) to visualize the proposed method, demonstrating that our method is able to locate expression-related regions more accurately than the baseline. In other words, the regions associated with each expression category that our model focuses on are more consistent with the regions humans understand as related to this expression. Grad-CAM can pinpoint the area that the deep neural network focuses on under a given category, making the decision-making process more interpretable and visible. The experimental results of some examples on three databases are demonstrated in Fig. 7. We show visualizations of categories

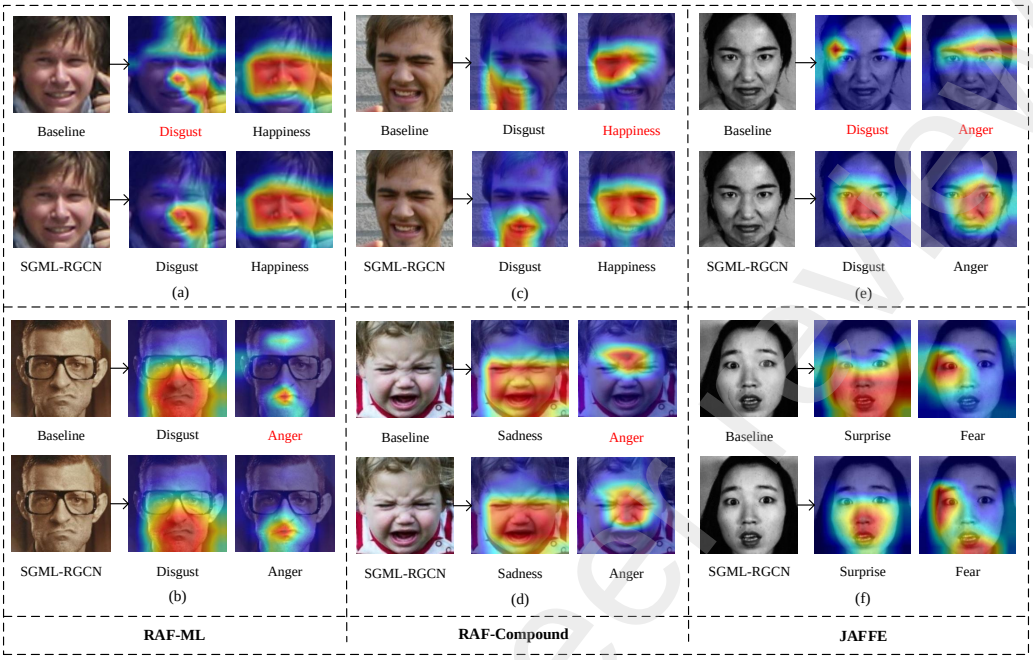

Figure 7: Visualization of baseline and proposed SVML-RGCN using Grad-CAM on the test sets of RAF-ML (top), RAF-Compound (middle) and JAFFE (bottom). The red labels below the images indicate the categories that the model failed to predict.

corresponding to images. For each group of images, the left column is the original image, while the middle and right columns are the heat maps generated by the Grad-CAM method. The labels are marked below the images, with the red labels indicating the categories that the model failed to predict.

Compared with the baseline model, our SVML-RGCN method demonstrates superior performance in identifying facial regions associated with expressions and effectively eliminates irrelevant factors such as hair. For instance, in group (a), the baseline model incorrectly focuses on hair when identifying "disgust". While the SVML-RGCN model focuses on the expression-related area, which is near the corner of the mouth. SVML-RGCN also pays attention to the eyes, which are also related to "disgust". In group (b), the baseline model mistakenly concerns about the blank part of the forehead when recognizing "anger" expression, while the SVML-RGCN model avoids this problem. Additionally, in group (c), SVML-RGCN surpasses the baseline model in accurately identifying both types of facial expressions. When locating "disgust", our model more accurately locates the area asso-

ciated with "baring teeth and grinning". And when identifying "happiness", SVML-RGCN also successfully locates the shape of eyes' region, together with the raised corners of the mouth. As for group (d), the baseline model only focuses on the area near the eyebrows when trying to identify "anger". What's worse, it pays more attention to the forehead with little information. SVML-RGCN, on the other hand, In contrast, comprehensively focuses on areas related to "anger", including the furrowed brows, eyes, and upper part of the nose. Then turning to group (e), we can find that when recognizing "disgust", the baseline model only pays attention to the tip of brow which has little useful information and the hair which is considered as interfering factor. In contrast, SVML-RGCN identifies the furrowed brow, wrinkled nose, and pursed lips that are indicative of the expression.

Even for conditions that both models classify correctly (e.g., group (f)), the localization of SVML-RGCN is more accurate compared to the baseline. Specifically, when "surprise" is considered in group (f), the baseline model pays extra attention to useless hair and background regions. When recognizing "fear", compared to the basline, SVML-RGCN not only focuses on widened eyes, but also on the open mouth, which is related to the expression as well.

5. Conclusion

In this paper, we propose the SVML-RGCN model, which utilizes the label correlations to improve the performance of ML-FER. Specifically, we apply R-GCN layers to model both positive and negative relations between different expression labels. Additionally, we use prior label representations, which combine semantic and visual information, as nodes of the graph. The semantic information comes from word embeddings of the labels, and the visual information is provided by visual features of basic expressions. Extensive quantitative experiments on three FER databases have proved the effectiveness of the proposed model. In addition, we demonstrate and analyze the interpretability of the model.

In future work, We will take a stab at using Transformer as backbone. Contrastive Language–Image Pre-training (CLIP) Radford et al. (2021) is a widely adopted vision language pre-trained model. We intend to utilize the image encoder of CLIP to obtain the features of expression images and the text encoder to provide the semantic information of different expressions.

Acknowledgments

This work was supported by the Science and Technology Innovation 2030 (STI2030)- Major Project (2021ZD0200600).

The authors would like to thank the associate editor and the reviewers for their comments and suggests for this paper.

References

Ben, X., Gong, C., Huang, T., Li, C., Yan, R., Li, Y., 2023. Tackling micro-expression data shortage via dataset alignment and active learning. IEEE Transactions on Multimedia 25, 5429–5443. doi:10.1109/TMM.2022.3192727.

Candra, H., Yuwono, M., Chai, R., Nguyen, H.T., Su, S., 2016. Classification of facial-emotion expression in the application of psychotherapy using viola-jones and edge-histogram of oriented gradient, in: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 423–426.

Chen, Z.M., Wei, X.S., Wang, P., Guo, Y., 2019. Multi-label image recognition with graph convolutional networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5177–5186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Du, L., Hu, H., 2019. Weighted patch-based manifold regularization dictionary pair learning model for facial expression recognition using iterative optimization classification strategy. Computer Vision and Image Understanding 186, 13–24. URL: https://www.sciencedirect.com/science/article/pii/S1077314219300943, doi:https://doi.org/10.1016/j.cviu.2019.06.003.

Ekman, P., 1984. Expression and the nature of emotion. Approaches to emotion 3, 344.

Fabian Benitez-Quiroz, C., Srinivasan, R., Martinez, A.M., 2016. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5562–5570.

Feng, S., Wan, H., Wang, N., Luo, M., 2021. Botrgcn: Twitter bot detection with relational graph convolutional networks, in: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 236–239.

Friesen, E., Ekman, P., 1978. Facial action coding system: a technique for the measurement of facial movement. Palo Alto 3, 5.

Gu, Y., Yan, H., Zhang, X., Wang, Y., Ji, Y., Ren, F., 2023. Toward facial expression recognition in the wild via noise-tolerant network. IEEE Transactions on Circuits and Systems for Video Technology 33, 2033–2047. doi:10.1109/TCSVT.2022.3220669.

Guo, H., Zheng, K., Fan, X., Yu, H., Wang, S., 2019. Visual attention consistency under image transforms for multi-label image classification, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 729–739. doi:10.1109/CVPR.2019.00082.

Hassin, R.R., Aviezer, H., Bentin, S., 2013. Inherently ambiguous: Facial expressions of emotions, in context. Emotion Review 5, 60–65.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Jiang, J., Deng, W., 2023. Improving multi-label facial expression recognition with consistent and distinct attentions. IEEE Transactions on Affective Computing .

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 .

Kollias, D., 2023. Multi-label compound expression recognition: C-expr database & network, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5589–5598. doi:10.1109/CVPR52729.2023.00541.

Li, S., Deng, W., 2019. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. International Journal of Computer Vision 127, 884–906.

Li, S., Deng, W., Du, J., 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2852–2861.

Li, W., Bao, R., Harimoto, K., Chen, D., Xu, J., Su, Q., 2021a. Modeling the stock relation with graph network for overnight stock movement prediction, in: Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence, pp. 4541–4547.

Li, W., Luo, M., Zhang, P., Huang, W., 2021b. A novel multi-feature joint learning ensemble framework for multi-label facial expression recognition. IEEE Access 9, 119766–119777.

Li, Y., Gao, Y., Chen, B., Zhang, Z., Lu, G., Zhang, D., 2021c. Self-supervised exclusive-inclusive interactive learning for multi-label facial expression recognition in the wild. IEEE Transactions on Circuits and Systems for Video Technology 32, 3190–3202.

Lin, T.Y., RoyChowdhury, A., Maji, S., 2015. Bilinear cnn models for fine-grained visual recognition, in: Proceedings of the IEEE international conference on computer vision, pp. 1449–1457.

Liu, H., Cai, H., Lin, Q., Li, X., Xiao, H., 2022. Adaptive multilayer perceptual attention network for facial expression recognition. IEEE Transactions on Circuits and Systems for Video Technology 32, 6253–6266. doi:10.1109/TCSVT.2022.3165321.

Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J., 1998. Coding facial expressions with gabor wavelets, in: Proceedings Third IEEE international conference on automatic face and gesture recognition, IEEE. pp. 200–205.

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543.

Plutchik, R., 1991. The emotions. University Press of America.

Pons, G., Masip, D., 2020. Multitask, multilabel, and multidomain learning with convolutional networks for emotion recognition. IEEE Transactions on Cybernetics 52, 4764–4771.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR. pp. 8748–8763.

Rosenberg, E.L., Ekman, P., 2005. What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Oxford University Press. URL: https://doi.org/10.1093/acprof:oso/9780195179644.001.0001, doi:10.1093/acprof:oso/9780195179644.001.0001.

Schlichtkrull, M., Kipf, T.N., Bloem, P., Van Den Berg, R., Titov, I., Welling, M., 2018. Modeling relational data with graph convolutional networks, in: The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15, Springer. pp. 593–607.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, pp. 618–626.

Solis-Arrazola, M.A., Sanchez-Yañez, R.E., Garcia-Capulin, C.H., Rostro-Gonzalez, H., 2024. Enhancing image-based facial expression recognition through muscle activation-based facial feature extraction. Computer Vision and Image Understanding 240, 103927. URL: https://www.sciencedirect.com/science/article/pii/S1077314224000080, doi:https://doi.org/10.1016/j.cviu.2024.103927.

Tallec, G., Yvinec, E., Dapogny, A., Bailly, K., 2022. Multi-label transformer for action unit detection. arXiv preprint arXiv:2203.12531 .

Teng, J., Zhang, D., Zou, W., Li, M., Lee, D.J., 2023. Typical facial expression network using a facial feature decoupler and spatial-temporal learning. IEEE Transactions on Affective Computing 14, 1125–1137. doi:10.1109/TAFFC.2021.3102245.

Wang, L., Zhang, X., Jiang, N., Wu, H., Yang, J., 2022. D 2 s: Dynamic distribution supervision for multi-label facial expression recognition, in: 2022 IEEE International Conference on Multimedia and Expo (ICME), IEEE. pp. 1–6.

Xu, Y., Yang, J., 2019. Look again at the syntax: Relational graph convolutional network for gendered ambiguous pronoun resolution. arXiv preprint arXiv:1905.08868 .

Yao, L., Mao, C., Luo, Y., 2019. Graph convolutional networks for text classification, in: Proceedings of the AAAI conference on artificial intelligence, pp. 7370–7377.

Yao, T., Pan, Y., Li, Y., Mei, T., 2018. Exploring visual relationship for image captioning, in: Proceedings of the European conference on computer vision (ECCV), pp. 684–699.

Yu, W., Xu, H., 2022. Co-attentive multi-task convolutional neural network for facial expression recognition. Pattern Recognition 123, 108401.

Yu, Z., Yu, J., Fan, J., Tao, D., 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in: Proceedings of the IEEE international conference on computer vision, pp. 1821–1830.

Zhang, J., Wang, W., Li, X., Han, Y., 2024. Recognizing facial expressions based on pyramid multi-head grid and spatial attention network. Computer Vision and Image Understanding 244, 104010. URL: https://www.sciencedirect.com/science/article/pii/S1077314224000912, doi:https://doi.org/10.1016/j.cviu.2024.104010.

Zhang, Y., Wang, C., Ling, X., Deng, W., 2022. Learn from all: Erasing attention consistency for noisy label facial expression recognition, in: European Conference on Computer Vision, Springer. pp. 418–434.

Zheng, W., Zong, Y., Zhou, X., Xin, M., 2018. Cross-domain color facial expression recognition using transductive transfer subspace learning. IEEE Transactions on Affective Computing 9, 21–37. doi:10.1109/TAFFC.2016.2563432.