MGRR-Net: Multi-level Graph Relational Reasoning Network for Facial Action Unit Detection

XURI GE, Unviersity of Glasgow, UK
JOEMON M. JOSE, Unviersity of Glasgow, UK
SONGPEI XU, Unviersity of Glasgow, UK
XIAO LIU, Tencent, China
HU HAN, Institute of Computing Technology, Chinese Academy of Sciences and University of the Chinese Academy of Sciences, China

The Facial Action Coding System (FACS) encodes the action units (AUs) in facial images, which has attracted extensive research attention due to its wide use in facial expression analysis. Many methods that perform well on automatic facial action unit (AU) detection primarily focus on modelling various AU relations between corresponding local muscle areas or mining global attention-aware facial features; however, they neglect the dynamic interactions among local-global features. We argue that encoding AU features just from one perspective may not capture the rich contextual information between regional and global face features, as well as the detailed variability across AUs, because of the diversity in expression and individual characteristics. In this paper, we propose a novel Multi-level Graph Relational Reasoning Network (termed *MGRR-Net*) for facial AU detection. Each layer of MGRR-Net performs a multi-level (*i.e.*, region-level, pixel-wise and channel-wise level) feature learning. On the one hand, the region-level feature learning from the local face patch features via graph neural network can encode the correlation across different AUs. On the other hand, pixel-wise and channel-wise feature learning via graph attention networks (GAT) enhance the discrimination ability of AU features by adaptively recalibrating feature responses of pixels and channels from global face features. The hierarchical fusion strategy combines features from the three levels with gated fusion cells to improve AU discriminative ability. Extensive experiments on DISFA and BP4D AU datasets show that the proposed approach achieves superior performance than the state-of-the-art methods.

CCS Concepts: • **Computing methodologies** → **Computer vision**; **Biometrics**; **Image representations**.

Additional Key Words and Phrases: Facial action units, graph attention network, local-global interaction, multi-level relational reasoning

1 INTRODUCTION

Facial action units (AUs) are defined as a set of facial muscle movements that correspond to a displayed expression according to the Facial Action Coding System(FACS) [8]. As a fundamental research problem, AU detection is beneficial to facial expression analysis [26, 65, 67], and has wide potential applications in diagnosing mental health issues [40, 48], improving e-learning experiences [37], detecting deception [22], *etc.* However, AU detection is challenging because of the difficulty

Authors' addresses: Xuri Ge, Unviersity of Glasgow, School of Computing Science, Glasgow, UK, x.ge.2@research.gla.ac.uk; Joemon M. Jose, Unviersity of Glasgow, School of Computing Science, Glasgow, UK, joemon.jose@glasgow.ac.uk; Songpei Xu, Unviersity of Glasgow, School of Computing Science, Glasgow, UK, s.xu.1@research.gla.ac.uk; Xiao Liu, Tencent, Online Media Business, Beijing, China, ender.liux@gmail.com; Hu Han, Institute of Computing Technology, Chinese Academy of Sciences and University of the Chinese Academy of Sciences, Beijing, China, hanhu@ict.ac.cn.

XXXX-XXXX/2024/2-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

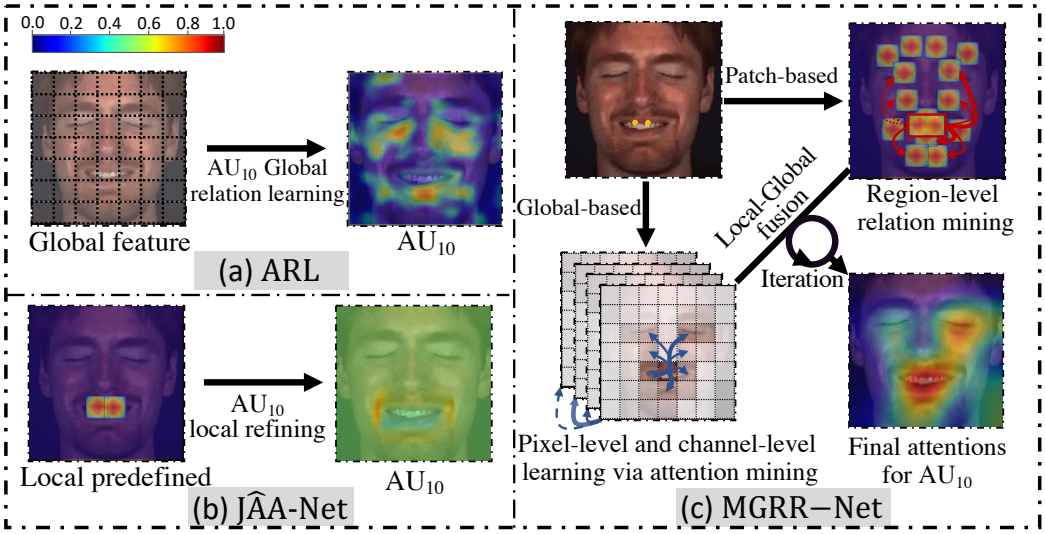

Fig. 1. Comparisons between the proposed method and two state-of-the-art methods in AU feature learning, and the corresponding visualized activation maps for AU10 (Upper Lip Raiser / Levator labii superioris). (a) ARL [45] performs global feature learning, (b) JÂANet [44] learns from predefined local regions based on the landmarks, and (c) multi-level feature learning from both local regions and global face regions (best viewed in color).

in identifying the subtle facial changes caused by AUs and individual physiology. Some earlier studies [25, 56] design hand-crafted features to represent different local facial regions related to AUs, according to the corresponding movements of facial muscles. However, hand-crafted shallow features are not discriminative enough to represent the rich facial morphology. Hence, deep learning-based AU detection methods that rely on global and local facial features have been studied to enhance the feature representation of each AU.

Several recent works [29, 36, 41, 45] aim to enhance the corresponding AU feature representation by combining the affected features in a deep global face feature map. For instance, LP-Net [36] using an LSTM model [13] combines the patch features from grids of equal partition made by a global Convolutional Neural Network (CNN). ARL [45] directly learns spatial attention from the global CNN features of independent AU branches, as shown in Fig. 1 (a). And [32] separately represented AU features directly from a shared full-face feature via multiple independent fully connected layers to model the relationships among all AUs in a graph. However, these methods suffered from the challenges of accurate localization of muscle areas corresponding to AUs, leading to potential interference from some irrelevant regions. In the past, such issues were addressed by extracting AU-related features from regions of interest (ROIs) centered around the associated facial landmarks [43, 44, 71], which provide more precise muscle locations for AUs and lead to a better AU detection performance. For example, JAA [43] and JÂANet [44] propose attention-based deep models to adaptively select the highly-contributing neighboring pixels of initially predefined muscle region for joint AU detection and face alignment, as shown in the Fig. 1 (b). However, the above local attention-based methods emphasize learning the appearance representation of each facial region based on detected landmarks while ignoring some intrinsic dependencies between different facial muscles. For example, AU2 (“Outer Brow Raiser") and AU7 (“Lid Tightener") will be activated simultaneously when scaring and AU6 (“Cheek Raiser") and AU12 (“Lip Corner Puller"), usually simultaneously

in a smiling face. To this end, some methods [6, 30, 35, 69] try to utilize prior knowledge of AU correlation by defining a fixed graph that represents the statistical AU correlations. For instance, [30] constructs a predefined graph for each face based on the AU co-occurrences to explicitly model the relationships between AU regions and enhance their semantic representations. However, it is difficult to effectively capture the dynamic relationships between AUs and the distinction of related AUs by a single predefined graph due to the complexity of AU activation and diversity across different subjects. Recent works [49–51] make an attempt to exploit an adaptive graph to model the uncertainty relationship between AUs. For instance, [50] emphasises the learning of important local facial regions based on probabilistic graph and obtain better facial appearance features by emphasizing important local facial regions via Long Short-Term Memory (LSTM) [11]. However, these approaches still enhance the semantic AU representations from the perspective of better regional feature representation, neglecting the modelling of the distinctive local and global features of each AU.

The key issue of facial AU detection lies in obtaining a better facial appearance representation by improving the feature discriminative ability of local AUs and global features from the whole face. On the one hand, region-level dynamic AU relevance mining based on facial landmarks accurately detects the corresponding muscles and flexibly models the relevance among muscle regions. It is different from the existing methods focusing on extracting features for a single AU region [43, 44] or a predefined fixed graph representing prior knowledge [19]. Although there have been many methods [30, 49–51] on modelling relationships between AU regions, this issue still needs to be addressed effectively. On the other hand, due to the differences in expressions, postures and individuals, fully learning the responses of the target AU in the global face can better capture the contextual differences between different AUs and complement more semantic details from the global face. For instance, [43, 44] simply concatenated the global features extracted from the whole face via CNNs with all local AU features for input into the final classifier. However, it is difficult for all these methods to learn the sensitivity of the target AU within the global face and supplement enough semantic details from the global face representation in terms of different expressions, postures and individuals. To the best of our knowledge, how to better respond globally to each AU remains unexploited in existing works [19, 30, 32, 44].

Motivated by the above insights, we propose a novel technique for facial AU detection called MGRR-Net. Our main innovations lie in three aspects, as shown in Fig. 1 (c). Firstly, we introduce a dynamic graph to model and reason the relationship between a target AU and other AUs. The region-level AU features (as nodes) can accurately locate the corresponding muscles. Secondly, we supplement each AU with different levels (channel- and pixel-level) of attention-aware details from global features, which greatly improves the distinction between AUs. Finally, we iteratively refine the AU features of the proposed multi-level local-global relational reasoning layer, which makes them more robust and more interpretable. Different from the existing GNN-based approaches [19, 30, 32, 35, 49, 51] that utilize complex GCNs [18] to enhance the distinguishability of AUs by constructing AU relationships, however, we supplement each AU with different perspectives (channel- and pixel-level) of attention-aware details from global features, making it possible to achieve the same purpose in a basic GNN and solve a certain over-smoothing issue. In particular, we extract the global features by multi-layer CNNs and precise AU region features based on the detected facial landmarks, which serve as the inputs of each multi-level relational reasoning layer. A simple region-level AU graph is constructed to represent the relationships by the adjacency matrix (as edges) among AU regions (as nodes), initialized by prior knowledge and iteratively updated. We propose a method to learn channel- and pixel-wise semantic relations for different AUs at the same time by processing them in two separate efficient and effective multi-head graph attention networks (MH-GATs) [58]. Through this, we model the complementary channel- and pixel-level global details.

After these local and global relation-oriented modules, a hierarchical gated fusion strategy helps to select more useful information for the final AU representation in terms of different individuals.

The contributions of this work are as follows:

- We propose a novel end-to-end iterative reasoning and training scheme for facial AU detection, which leverages the complementary multi-level local-global feature relationships to improve the robustness and discrimination for AU detection;
- We construct a region-level AU graph with the prior knowledge initialization and dynamically reason the correlated relationship of individual AUs, thereby improving the robustness of AU detection;
- We propose a GAT-based model to improve the discrimination of each local AU patch by supplementing multiple levels of global features;
- The proposed MGRR-Net outperforms the state-of-the-art approaches for AU detection on two widely used benchmarks, *i.e.,* BP4D and DISFA, without any external data or pre-trained models.

2 RELATED WORK

2.1 Facial AU Detection

Automatic AU detection has been studied for decades, and several methods [23, 33, 36, 44, 45, 66, 71] have been proposed. Some works [23, 29, 36, 41, 45] predicted the activation state of each AU by directly extracting global face features via CNNs. For instance, [23, 45] proposed sequential or parallel channel and spatial attention learning mechanisms to explore the attention-aware global representation of each face. While progress has indeed been achieved through the utilization of global representations, the advancement remains constrained by the rudimentary nature of the coarse-grained features. Most existing approaches for facial AU detection use feature learning from local patches [20, 33, 36, 43, 44, 71]. However, there is a need to pre-define the patch location first in some early works [53, 70]. For instance, [15] proposed to use domain knowledge and facial geometry to pre-select a relevant image region (as a patch) for a particular AU and feed it to a convolutional and bi-directional Long Short-Term Memory (LSTM) [11] neural network. [43] proposed an end-to-end deep learning framework for joint AU detection and face alignment, which used the detected landmarks to locate specific AU regions. However, all the above methods focused only on independent regions without considering the correlations among different AU areas to reinforce and diversify each other. Recent works focus on capturing the relations among AUs for local feature enhancement, which can improve robustness compared to single-patch features or global face features. [19] incorporated the AU knowledge graph as extra guidance for enhancing facial region representation. [30] applied the spectral perspective of graph convolutional network (GCN) for AU relation modelling, which also needed an additional AU correlation reference extracted from EAC-Net [21]. However, these methods need prior knowledge of co-occurrence probability in different datasets to construct the fixed relation matrix instead of dynamically updating for different expressions and individuals. [10] proposed a complex skip-BiLSTM to mine the potential mutual assistance and exclusion relationship between AU branches and simple complementary global information. [51] proposed a performance-driven Monte Carlo Markov Chain to generate graphs from the global face, which, however, also captures some irrelevant regions affecting the performance. Moreover, these approaches usually ignored or simply fused the local and global information for each AU without considering the importance (important and non-important) of features. Recently, [32] learned a unique AU graph to explicitly describe the relationship between AUs, where each AU is simply represented from the same full face representation via a fully connected layer and a global average pooling. Although this method explores the global face features to some extent, it relies on strong global feature

extraction benchmarks and lacks accurate localization of local muscle areas and discriminable feature representation via local-global interaction.

2.2 Graph Neural Network

Integrating graphs with deep neural networks have recently been an emerging topic in deep learning research. GCNs have been widely used in many applications such as human action recognition [62], emotion recognition [52], social relationship understanding [60] and object parsing [27]. [19] proposed to apply a gated graph neural network (GGNN) with the guidance of AU knowledge-graph on facial AU detection. [35] embedded the relations among AUs through a predefined GCN to enhance the local semantic representation. However, these AU detection methods require a fixed predefined graph from different datasets when applying GGNN or GCN. [49, 51] applied an adaptive graph to model the relationships between AUs based on global features, ignoring local-global interactions. Recently, a novel graph attention network with multi-head (MH-GAT) leverages masked self-attentional layers to operate on graph-structured data, which shows high computational efficiency.

As far as we know, there has been no work attempting to obtain better feature representation by multiple interactions between local AU regions and the global face, which we believe is an important cue to boost facial AU detection performance with more fine-grained information and higher diversity of expressions. To this end, our proposed MGRR-Net automatically models the relevance among the facial AU regions by a dynamic matrix as a graph and supplements each AU patch with multiple levels of global features to improve the variability. Multiple layers of iterative refinement significantly improve the AU discrimination ability. Our MGRR-Net has wide potential applications in diagnosing mental health issues [40, 48], improving e-learning experiences [37], detecting deception [22], *etc.* For example, in our future work, we will apply MGRR-Net to automatically estimate facial palsy severity for patients, such as [9]. This will be helpful for the diagnosis and treatment of people who have facial palsy across the world.

3 APPROACH

As shown in Fig. 2, the proposed approach consists of two core modules in each relational reasoning layer, *i.e.*, region-level local feature learning with relational modelling, and global feature learning with channel- and pixel-level attention. A hierarchical gated fusion network is designed to combine multi-level local and global features as the new target AU feature. Finally, after multiple layers of iterative refinement and updating, the AU features are fed into a multi-branch classification network for AU detection. For clarity, the main notations and their definitions throughout the paper are shown in Table 1.

3.1 Global and Local Features Extraction

Given a face image I, we adapt a stem network from the widely used multi-branch network [43] to extract the original global feature O_G and further obtain the AU regions based on the detected landmarks. Different from the [21], our stem network contains a face alignment module for automatic face landmark detection, facilitating end-to-end training of our method. All branches share the stem network to reduce training costs and the complexity of network training. In particular, a hierarchical and multi-scale region learning module in the stem network extracts features from each local patch with different scales, thus obtaining multi-scale representations. A series of landmarks $S = \{s_1, s_2, ..., s_m\}$ with length m are detected by an efficient face alignment module similar to [44], including three convolutional blocks connected to a max-pooling layer. According to the detected landmarks, local patches are calculated, and their features $V = \{v_1, v_2, ..., v_n\}$ are learned via the stem network, where n is the number of selected AU patches. For simplicity, we do not repeat the detailed structure of the stem network here.

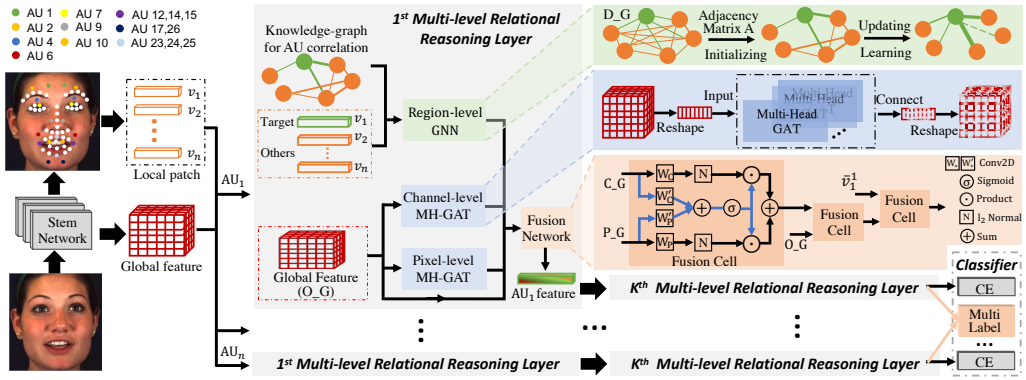

Fig. 2. The overall architecture of the proposed MGRR-Net for facial AU detection. Given one face image, the region-level features of local AU patches are extracted based on the detected landmarks from an efficient landmark localization network. The original global feature is extracted from the same shared stem network. Then the region-level GNN initialized with prior knowledge is applied to encode the correlation between different AU patches. Two separate MH-GATs are adopted to get two levels of global attention-aware features to supplement each AU. Finally, multiple levels of local-global features are fused by a hierarchical gated fusion strategy and refined by multiple iterations (best viewed in color).

3.2 Multi-level Relational Reasoning Layer

After we get the original global feature O_G for a face and the local region features $V = \{v_1, v_2, ..., v_n\}$ for AUs, a multi-layer multi-level relational reasoning model is introduced to automatically explore the relationship of individual local facial regions and supply two levels of global information. Fig. 2 shows the detailed structure of the 1^{st} multi-level relational reasoning layer.

3.2.1 Region-level Local Feature Relational Modeling. Different from the predefined fixed AU relationship graph in [19], we construct a fully-connected graph D_G for all AUs, where the region-level features $V = \{v_1, v_2, ..., v_n\}$ constitute the nodes, and a learnable adjacency matrix A constitutes the edges at each layer to represent the possibility of AU co-occurrence (co-activated or non-activated). In this scheme, the AUs with no co-occurrence or low co-occurrence relationship in the training set will not be completely ignored, like [19, 30, 35]. During the training process, we utilize prior knowledge to initialize A to assist and constrain model learning. Specifically, the dynamic graph D_G comprises nodes (the local region features $V = \{v_1, v_2, ..., v_n\}$) and edges (the relationship matrix A among AUs). Following [35], we calculate the relationship coefficients between AUs from datasets to initialize the adjacency matrix A (Fig. 5 (a) shows the predefined AU correlation on BP4D). The statistical prior knowledge serves as the initial relationship, allowing suppression of the edges with low correlation and speeding up the relationship learning. The relationship coefficient A_{ij} between the i-th and j-th AU can be formulated as:

$$P_{ij} = \frac{1}{2}(P(a_i = 1|a_j = 1) + P(a_i = 0|a_j = 0)), \tag{1}$$

$$A_{ij} = |(P_{ij} - 0.5) * 2| \tag{2}$$

where a_i=1 denotes i-th AU is activated and 0 otherwise, $|\cdot|$ means absolute value function. From Eq.1 and Eq.2, $P(a_i$=1$|a_j$=1)=0.5 means that when j-th AU is activated, the probability of occurrence

Table 1. Main notations and their definitions.

Notation	Definition
I	a facial image
S	a set of detected landmarks
O_G	the original global feature
m	the number of detected landmarks
V	a set of calculated patch features
v_i	the feature of i-th patch
n	the number of calculated patches corresponding to AUs
D_G	a fully-connected graph for AU relationship construction
A	a learnable adjacency matrix
a_i	the activation status of the i-th AU
P_{ij}	the coefficient between i-th and j-th AU
P, C	a set of pixel- and channel-level features
P_G	the pixel-level attention-aware global feature
C_G	the channel-level attention-aware global feature
L	the number of parallel attention layers
K	the number of relational reasoning layers
$\bar{v}_i^k$	the feature of i-th AU patch after k-th reasoning layer
GFC	a gated fusion cell
(x_i, y_i)	the ground-truth coordinate of the i-th facial landmark
$(\hat{x}_i, \hat{y}_i)$	the predicted coordinate of the i-th facial landmark
d_o	the ground-truth inter-ocular distance
p_i	the ground-truth occurrence probability of i-th AU
$\hat{p}_i$	the predicted occurrence probability of i-th AU

is equal to the no occurrence for i-th AU. It indicates that the activation of j-th AU could not provide useful information for the i-th AU, and therefore no edge is connected.

3.2.2 Attention-aware Global Features Learning. We argue that complementary global feature can improve the discrimination between AUs, which also alleviates the over-smoothing issue in graph neural networks for local relationship modelling. To this end, we employ two separate high-efficiency GAT models [58] to perform channel- and pixel-level attention-aware global features from original deep visual features in order to handle expression and subject diversities. Specifically, we reshape the original global feature O_G $\in \mathbb{R}^{(c,w,h)}$ into a set of channel-level features $\{C_1, ..., C_c\}, C_i \in \mathbb{R}^{w*h}$. Similarly, by reshaping pixel dimensions and keeping channel dimension of O_G from a convolution layer to reduce the parameters, we get a set of pixel-level features $\{P_1, ..., P_{w'*h'}\}, P_i \in \mathbb{R}^c$. The attention coefficient α_{ij} between channel- or pixel-level features is calculated in GAT, which can be formulated as (Here we take the process of channel-level attention-aware features as an example.):

$$\alpha_{ij} = \frac{\exp(U_q C_i (U_k C_j)^T / \sqrt{D})}{\sum_{o \in \Omega_i} \exp(U_q C_i (U_o C_o)^T / \sqrt{D})}, \tag{3}$$

where U_q, U_k, U_o are the parameters of mapping from $w * h$ to D and Ω_i denotes neighborhoods of C_i. $\sqrt{D}$ acts as a normalization factor. Following [57, 58], we also employ multi-head dot product by L

parallel attention layers to speed up the calculation efficiency. The overall working flow is formulated as:

$$\begin{aligned} \bar{C}_i &= \mathrm{ReLU}(\sum\nolimits_{o\in\Omega_i} U_c||_l^L(\alpha_{io}^l * C_i)), \\ \alpha_{ij}^l &= \frac{\exp(U_q' C_i (U_k' C_j)^T/\sqrt{d})}{\sum_{o\in\Omega_i}\exp(U_q' C_i (U_o' C_o)^T/\sqrt{d})}, \end{aligned} \tag{4}$$

where U_c is the mapping parameter, U_q', U_k', U_o' map the feature dimension to $1/L$ of the original, $||$ means concatenation, and d equals D/L. Finally, the new channel-level attention-aware global feature C_G=$\{\bar{C}_i\}$ is reshaped to the same domination with O_G. With the same process on pixel-level features $\{P_1, ..., P_{w'*h'}\}$, we can get the final pixel-level attention-aware global features P_G after a deconvolution layer behind of a GAT with multi-head (MH-GAT).

3.2.3 Hierarchical Fusion and Iteration. We iteratively refine the i-th target AU feature of the proposed multi-level relational reasoning layer K times, which obtains other correlated local, and regional information and provides rich global details in each layer. The process can be formulated as:

$$\bar{v}_i^k = W_i^k v_i^k + \sum\nolimits_i^n (A_{ij}^k W_j^k v_j), \tag{5}$$

where W^k is the mapping parameter and A_{ij}^k means the learnable correlation coefficient between AU_i and AU_j at k-th layer. We then use a hierarchical fusion strategy by a gated fusion cell (GFC) to complement the global multi-level information for each updated AU feature at k-th layer as follows:

$$\bar{v}_i^{k+1} = \mathrm{GFC}(\bar{v}_i^k, \mathrm{GFC}(\mathrm{O\_G}^k, \mathrm{GFC}(\mathrm{C\_G}^k, \mathrm{P\_G}^k))), \tag{6}$$

We define the operation of GFC as follows:

$$\mathrm{GFC}(\mathrm{C\_G}^k, \mathrm{P\_G}^k) = \beta \odot \|W_C^k \mathrm{C\_G}^k\|_2 + (1-\beta) \odot \|W_P^k \mathrm{P\_G}^k\|_2, \tag{7}$$

$$\beta = \sigma(W_C^{k'} \mathrm{C\_G}^k + W_P^{k'} \mathrm{P\_G}^k), \tag{8}$$

where σ is the sigmoid function, and $\|\cdot\|$ denotes the l_2-normalization. $W_*^{k'}$ and W_*^k denote the Conv2D operation.

3.3 Joint Learning

A multi-label binary classifier is used to classify the AU activation state, which adopts a weighted multi-label cross-entropy loss function (denoted as CE in Fig. 2) as follows,

$$\mathcal{L}_{au} = -\frac{1}{n}\sum_{i=1}^{n} w_i[p_i \log \hat{p}_i + (1-p_i)\log(1-\hat{p}_i)], \tag{9}$$

where p_i and $\hat{p}_i$ denote the ground-truth and predicted occurrence probability of the i-th AU, respectively; w_i is the data balance weights used in [43]. Furthermore, we also minimize the loss of AU category classification $\mathcal{L}_{int}$ by integrating all AUs information, including the refined AU features and the face alignment features, which is similar to the processing of $\mathcal{L}_{au}$.

We jointly integrate face alignment and facial AU recognition into an end-to-end learning model. The face alignment loss is defined as:

$$\mathcal{L}_{align} = \frac{1}{2d_o^2}\sum_{i=1}^{m}[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2], \tag{10}$$

where (x_i, y_i) and $(\hat{x}_i, \hat{y}_i)$ denote the ground-truth coordinate and corresponding predicted coordinate of the i-th facial landmark, and d_o is the ground-truth inter-ocular distance for normalization [44].

Finally, the joint loss of our MGRR-Net is defined as:

$$\mathcal{L} = (\mathcal{L}_{au} + \mathcal{L}_{int}) + \lambda \mathcal{L}_{align}. \tag{11}$$

where λ is a tuning parameter for balancing.

4 EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the proposed MGRR-Net. Especially the dataset and training strategy are first introduced. Then, MGRR-Net is compared with state-of-the-art FAU detection approaches quantitatively. Finally, we qualitatively analyze the results in detail.

4.1 Dataset

We provide evaluations on the popular BP4D [68] and DISFA [34] datasets.

BP4D is a spontaneous facial AU database containing 328 facial videos from 41 participants (23 females and 18 males) who were involved in 8 sessions. Similar to [20, 44, 45], we consider 12 AUs and 140K valid frames with labels.

DISFA consists of 27 participants (12 females and 15 males). Each participant has a video of $4,845$ frames. We limited the number of AUs to 8, similar to [20, 44]. Following [43, 44], frames in DISFA with AU intensity labels higher than two are considered positive samples. Compared to BP4D, the experimental protocol and lighting conditions deliver DISFA to be a more challenging dataset.

During training, each frame of BP4D and DISFA is annotated with 49 landmarks detected and calculated by SDM [61]. Following the experiment setting of [43, 44], we evaluated the model using the 3-fold subject-exclusive cross-validation protocol.

4.2 Training Strategy

Our model is trained on a single NVIDIA RTX 2080Ti with 11 GB memory. The whole network is trained with the default initializer of PyTorch [39] with the SGD solver, a Nesterov momentum of 0.9 and a weight decay of 0.0005. The learning rate is set to 0.01 initially, with a decay rate of 0.5 every two epochs. The maximum epoch number is set to 15. During the training process, aligned faces are randomly cropped into 176×176 and horizontally flipped. Regarding the face alignment network and stem network, we set the value of the general parameters to be the same with [44]. The iteration layer number K is set to 2 except otherwise noted. The dimensionality of O_G is $(64, 44, 44)$ and D is 1024. We employ L=8 parallel attention layers in GATs. In our paper, all the mapping Conv2D operations used 1×1 convolutional filters with a stride one and a padding 1. We use a 3×3 Conv2D operation with a stride two and padding one before learning the channel-level feature to reduce the parameters. λ is empirically set to 0.5 for the joint optimisation of face alignment and facial AU detection on two benchmarks. Following the settings in [21, 44, 72], our MGRR-Net initializes the parameters of the well-trained model trained on BP4D when training on DISFA. This initialization greatly alleviates the poor performance issue on DISFA due to data volume and AU category imbalance. Compared to JÂA-Net [44], which takes 26.6ms per image to do a forward pass, our model takes just 16.5ms using an RTX 2080Ti GPU. This is due to the multi-head operation of the effective MH-GATs and the optimization of the model, which significantly reduces forward pass time. The training time is approximately 1.5 hours per epoch. In addition, we average the predicted probability of the local information and the integrated information as the final predicted activation probability for each AU rather than simply using the integrated information of all the AUs.

Table 2. Comparisons of AU recognition for 8 AUs on DISFA in terms of F1-frame score (in %). CLP† is a semi-supervised method. * means the method employed a pre-trained model on the additional dataset, such as ImageNet [7] and VGGFace2 [1], *etc.*

Method	AU Index								Avg.
	1	2	4	6	9	12	25	26	
DSIN [4]	42.4	39.0	68.4	28.6	46.8	70.8	90.4	42.2	53.6
JAA [43]	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
LP-Net [36]	29.9	24.7	72.7	46.8	49.6	72.9	93.8	65.0	56.9
ARL [45]	43.9	42.1	63.6	41.8	40.0	**76.2**	**95.2**	66.8	58.7
SRERL [19]	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
JÂAANet [44]	**62.4**	60.7	67.1	41.1	45.1	73.5	90.9	67.4	63.5
JAA-DGCN [16]	61.8	51.7	64.5	46.0	54.2	63.6	85.5	69.4	62.0
CLP† [24]	42.4	38.7	63.5	59.7	38.9	73.0	85.0	58.1	57.4
MMA-Net [42]	63.8	54.8	73.6	39.2	**61.5**	73.1	92.3	70.5	66.0
MGRR-Net	61.3	**62.9**	**75.8**	**48.7**	53.8	75.5	94.3	**73.1**	**68.2**
UGN-B* [49]	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
HMP-PS* [51]	21.8	48.5	53.6	56.0	58.7	57.4	55.9	56.9	61.0
DML* [59]	62.9	65.8	71.3	51.4	45.9	76.0	92.1	50.2	64.4
PIAP* [55]	50.2	51.8	71.9	50.6	54.5	79.7	94.1	57.2	63.8
TransAU* [14]	46.1	48.6	72.8	56.7	50.0	72.1	90.8	55.4	61.5
Bio-AU* [5]	41.5	44.9	60.3	51.5	50.3	70.4	91.3	55.3	58.2
MGRR-Net	61.3	62.9	**75.8**	48.7	53.8	75.5	94.3	**73.1**	**68.2**

Table 3. Comparisons of AU recognition for 8 AUs on DISFA in terms of Accuracy and AUC (in %). * means the method employed pretrained model on additional dataset.

AU	Accuracy						AUC				
	JAA [43]	ARL [45]	JÂAANet	MMA-Net [42]	UGN-B* [49]	**MGRR-Net**	DRML [72]	SRERL [19]	DML* [59]	DAR-GCN [17]	**MGRR-Net**
1	93.4	92.1	**97.0**	96.8	95.1	96.8	53.3	76.2	**90.5**	84.5	89.5
2	96.1	92.7	97.3	96.5	93.2	**97.4**	53.2	80.9	92.7	92.5	**93.0**
4	86.9	88.5	88.0	91.6	88.5	**92.7**	60.0	79.1	**93.8**	72.2	93.6
6	91.4	91.6	92.1	91.5	**93.2**	92.1	54.9	80.4	90.3	48.3	**91.1**
9	95.8	95.9	95.6	96.5	96.8	**96.9**	51.5	76.5	84.4	78.3	**91.9**
12	91.2	**93.9**	92.3	92.3	93.4	93.4	54.6	87.9	95.7	37.8	**95.9**
25	93.4	**97.3**	94.9	95.5	94.8	96.8	45.6	90.9	98.2	50.3	**99.0**
26	93.2	94.3	94.8	95.0	93.8	**95.6**	45.3	73.4	87.4	74.3	**94.4**
Avg.	92.7	93.3	94.0	94.5	93.4	**95.2**	52.3	80.7	91.6	67.3	**93.6**

Table 4. Comparisons with state-of-the-art methods for 12 AUs on BP4D in terms of F1-frame (in %). * means the method employed pretrained model on additional dataset.

AU	F1-frame															
	MLCR [35]	JAA [43]	LP-Net [36]	ARL [45]	SRERL[19]	JÂANet [44]	CLP [24]	MMA-Net [42]	**Ours**	R-CNN* [33]	UGN-B* [49]	HMP-PS*[51]	DML*[59]	TransAU* [14]	Bio-AU* [5]	**Ours**
1	42.4	47.2	43.3	45.8	46.9	**53.8**	47.7	52.5	[52.6]	50.2	54.2	53.1	52.6	51.7	57.4	[52.6]
2	36.9	44.0	38.0	39.8	45.3	47.8	**50.9**	**50.9**	[47.9]	43.7	46.4	46.1	44.9	49.3	52.6	[47.9]
4	48.1	54.9	54.2	55.1	55.6	58.2	49.5	**58.3**	[57.3]	57.0	56.8	56.0	56.2	61.0	64.6	[57.3]
6	77.5	77.5	77.1	75.7	77.1	78.5	75.8	76.3	[**78.5**]	78.5	76.2	76.5	79.8	77.8	79.3	[78.5]
7	77.6	74.6	76.7	77.2	78.4	75.8	**78.7**	75.7	[77.6]	78.5	76.7	76.9	80.4	79.5	81.5	[77.6]
10	83.6	84.0	83.8	82.3	83.5	82.7	80.2	83.8	[**84.9**]	82.6	82.4	82.1	85.2	82.9	82.7	[84.9]
12	85.8	86.5	87.2	86.6	87.6	88.2	84.1	87.9	[**88.4**]	87.0	86.1	86.4	88.3	86.3	85.6	[**88.4**]
14	61.0	61.9	63.6	58.8	63.9	63.7	67.1	63.8	[**67.8**]	67.7	64.7	64.8	65.6	67.6	67.8	[**67.8**]
15	43.7	43.6	45.3	47.6	**52.2**	43.3	52.0	48.7	[47.6]	49.1	51.2	51.5	51.7	51.9	47.3	[47.6]
17	63.2	60.3	60.5	62.1	**63.9**	61.8	62.7	61.7	[63.3]	62.4	63.1	63.0	59.4	63.0	58.0	[**63.3**]
23	42.1	42.7	48.1	47.4	47.1	45.6	45.7	46.5	[**47.4**]	50.4	48.5	49.9	47.3	43.7	47.0	[47.4]
24	55.6	41.9	54.2	**55.4**	53.3	49.9	54.8	54.4	[51.3]	49.3	53.6	54.5	49.2	56.3	44.9	[51.3]
Avg.	59.8	60.0	61.0	61.1	62.9	62.4	62.4	63.4	[**63.7**]	62.6	63.3	63.4	63.4	64.2	64.1	[63.7]

4.3 Evaluation Metrics

For all methods, the frame-based F1 score (F1-frame, %) is reported, which is the harmonic mean of the Precision P and Recall R and calculated by $F1 = 2P*R/(P+R)$. To conduct a more comprehensive comparison with other methods, we also evaluate the performance with AUC (%) refers to the area under the ROC curve and accuracy (%). In addition, the average results over all AUs (denoted as **Avg.**) are computed with "%" omitted.

4.4 Comparison with State-of-the-art Methods

We compare our proposed MGRR-Net with several frame-based AU detection baselines and the latest state-of-the-art methods, including Deep Structure Inference Network (DSIN) [4], Joint AU Detection and Face Alignment (JAA) [43], Multi-Label Co-Regularization (MLCR) [35], Local relationship learning with Person-specific shape regularization (LP-Net) [36], Attention and Relation Learning (ARL) [45], Semantic Relationships Embedded Representation Learning (SRERL) [19], Joint AU detection and face alignment via Adaptive Attention Network (JÂANet) [44], Data-Aware Relation Graph Convolutional Neural network (DAR-GCN) [17], Dual-channel Graph Convolutional Neural Network (JAA-DGCN) [16], a semi-supervised Contrastively Learning the Person-independent representations method (CLP) [24] and a Multiview Mixed Attention based Network (MMA-Net) [42]. To ensure reliable and fair comparisons, we directly use the results of these methods reported. Note that, the best and second-best results are shown using bold and underline, respectively. The experimental results of our MGRR-Net are shown with a grey background.

For a more comprehensive display, we present methods (marked with $*$) [2, 3, 5, 14, 49, 51, 55, 59] that use additional data, such as ImageNet [7] and VGGFace2 [1], for pre-training their complex feature extraction stem network firstly, such as ResNet [12] *etc*. From [14, 36], the pre-trained feature

Table 5. Comparisons with state-of-the-art methods for 12 AUs on BP4D in terms of Accuracy and AUC respectively (in %). * means the method employed pretrained model on additional dataset, such as ImageNet [7], *etc.* So we do not directly compare.

AU	Accuracy					AUC			
	UGN-B* [49]	JAA [43]	ARL [45]	JÂANet [44]	**MGRR-Net**	DRML [72]	SRERL [19]	DML* [59]	**MGRR-Net**
1	78.6	74.7	73.9	75.2	**78.7**	55.7	67.6	**78.5**	78.1
2	80.2	80.8	76.7	80.2	**82.1**	54.5	70.0	75.9	**77.2**
4	80.0	80.4	80.9	**82.9**	81.6	58.8	73.4	**84.4**	83.8
6	76.6	78.9	78.2	**79.8**	78.7	56.6	78.4	**88.6**	88.4
7	72.3	71.0	**74.4**	72.3	73.7	61.0	76.1	**84.8**	82.3
10	77.8	80.2	79.1	78.2	**81.2**	53.6	80.0	**87.3**	86.3
12	84.2	85.4	85.5	86.6	**86.9**	60.8	85.9	**93.9**	93.6
14	63.8	64.8	62.8	65.1	**67.0**	57.0	64.4	71.8	**72.9**
15	84.0	83.1	84.7	81.0	**84.2**	56.2	75.1	80.7	**80.8**
17	72.8	73.5	**74.1**	72.8	72.2	50.0	71.7	75.0	**78.2**
23	82.8	82.3	82.9	82.9	**84.1**	53.9	71.6	78.7	**79.3**
24	86.4	85.4	85.7	**86.3**	86.0	53.9	74.6	84.3	**87.8**
Avg.	78.2	78.4	78.2	78.6	**79.7**	56.0	74.1	82.0	**82.4**

extractor improved the average F1-score by at least 1.2% on BP4D. Due to the fact that our stem network only consists of a few simple convolutional layers, even if we pre-trained on additional datasets, it is unsuitable compared to pre-training on deeper feature extraction networks, such as ResNet50 [12], ResNet101 [12] and Swin Transformer-Base [31]. To this end, we have grouped them together to facilitate comparison with our proposed MGRR-Net. Notably, our results show excellence, affirming the superiority and efficacy of our proposed learning methodology. To provide a fair comparison, we omit the need for additional modality inputs and non-frame-based models [28, 47, 54, 63, 64].

4.4.1 Quantitative Comparison on DISFA. We compare our proposed method with its counterpart in Table 2 and Table 3. It has been shown that our MGRR-Net outperforms all its competitors with impressive margins. Compared with the existing end-to-end feature learning and multi-label classification methods DSIN [4] and ARL [45], our MGRR-Net shows significant improvements on all AUs. These results demonstrate the effectiveness of accurate muscle region localization for AU detection. Although ARL [45] also performs sequential multiple attention explorations on global features, we believe that the sequential mechanism may destroy the diversity of different attention-aware features and slow down the training time. JÂANet is the latest state-of-the-art method which also joint AU detection and face alignment into an end-to-end multi-label multi-branch network. Compared with the baseline JÂANet [44], our MGRR-Net increases the average F1-frame and average accuracy scores by large margins of 4.7% and 1.2% and shows clear improvements for most annotated AU categories. The main reason lies in JÂANet [44] completely ignores the correlation between branches and the individual modelling of each AU. Compared with JAA-DGCN [16] that also applies the graph relationship model, our MGRR-Net still performs better on most metrics

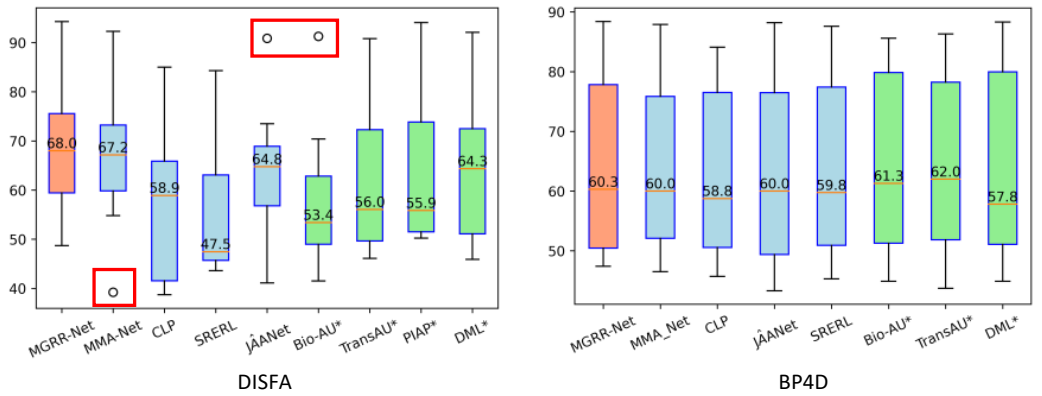

Fig. 3. Box Plots of the distribution of performances on all AU categories (the labeled values are medians). (a) on DISFA 3-flod test set and (b) on BP4D 3-flod test set.

because we model local relationships while supplementing a variety of information from the global face. Moreover, compared with the latest state-of-the-art MMA-Net [42], MGRR-Net achieves a 2.2% lead in the average F1-frame metric. In addition, compared with the current state-of-the-art AU detection methods based on pre-trained models, such as UGN-B [49], HMP-PS [51], DML [59], PIAP [55] and Bio-AU [5] *etc.*, we also achieve the best performance in terms of the average F1-frame.

Furthermore, the results of the Accuracy and AUC evaluations provide further evidence of the effectiveness of our method compared to other state-of-the-art methods. In particular, our MGRR-Net obtains a significant improvement on the average of Accuray, *i.e.* 95.2 % *vs.* 94.5%, compared with MMA-Net [42]. And on AUC metric, our MGRR-Net also achieves higher results on most metrics and increases 2.0% compared to DML* [59].

4.4.2 Quantitative Comparison on BP4D. Table 4 and 5 show the AU detection results of different methods in terms of F1-frame, Accuracy and AUC on BP4D dataset, where the method in the left of Table 4 uses a feature extractor without pre-training and the method with * is based on the pre-trained feature extractor (our method is trained on BP4D only). Compared with the multi-branch combination-based JÂANet [44], the average F1 frame score and average accuracy score of MGRR-Net get 1.3% and 1.1% higher, respectively. Furthermore, compared with the latest graph-based relational modelling method SRERL [19], MGRR-Net increases the average F1-frame and average AUC by large margins of 0.8% and 8.3%. This is mainly due to the fact that the proposed method models the semantic relationships among AUs while also gaining complementary features from multiple global perspectives to increase the distinguishability of each AU. In addition, our MGRR-Net achieves the best or second-best AU detection performance in terms of F1-frame, Accuracy and AUC for most of the 12 AUs annotated in BP4D compared with the state-of-the-art methods. For example, compared with the latest method MMA-Net [42], which simultaneously modelled the deep feature learning and the structured AU relationship in a unified framework, ours greatly outperforms it by 0.3% in terms of the average of F1-frame. In addition, compared with the advanced models pre-trained with additional data (marked with $*$ in Table 4 and Table 5), our MGRR-Net still has strong competitiveness.

Experimental results of MGRR-Net demonstrate its effectiveness in improving AU detection accuracy on DISFA and BP4D, as well as good robustness and generalization ability. Note that

Table 6. Effectiveness of key components of MGRR-Net evaluated on DISFA in terms of F1-frame score (in %).

Method		1	2	3	4	5	6	MGRR-Net
Setting	D_G	-	√	√	√	√	√	√
	O_G	-	-	√	-	√	√	√
	C_G	-	-	-	√	√	-	√
	P_G	-	-	-	√	-	√	√
AU Index	1	47.1	52.5	58.4	60.0	65.4	61.0	[61.3]
	2	61.1	58.1	63.0	65.7	64.5	67.3	[62.9]
	4	66.3	73.3	70.9	67.4	72.5	76.8	[75.8]
	6	44.7	44.4	46.2	43.8	42.6	40.9	[48.7]
	9	52.2	52.5	47.7	57.1	52.9	58.0	[53.8]
	12	74.9	73.2	72.1	75.4	75.3	74.8	[75.5]
	25	92.2	94.7	93.4	93.3	94.3	93.7	[94.3]
	26	66.2	71.2	71.8	64.7	71.4	65.8	[73.1]
Avg.		63.1	65.0	65.4	65.9	67.4	67.3	[68.2]

the main reason why some AUs are clearly less accurate than others is due to data imbalance, as shown in Figure 3, this is a phenomenon that exists in all existing methods [5, 14, 24, 42, 44, 55]. In BP4D, where the data distribution is relatively reasonable, the results' distribution of each method is close. But in DISFA, where the data distribution is more extreme, the result distribution of our MGRR-Net can perform better, *i.e.* lower variance and no outliers. We infer that two aspects promote this improvement. On one hand, we use a weighted multi-label cross-entropy loss function as Eq.(9) to solve the data imbalance problem to a certain extent. On the other hand, our multi-level fused representation can complement each AU representation, as well as combine with other AU areas, to further improve AU classification.

4.5 Ablation Studies

We perform detailed ablation studies on DISFA to investigate the effectiveness of each part of our proposed MGRR-Net. Due to space limitations, we do not show the ablation results for BP4D, but it is consistent with DISFA. To assess the effect of different components, we run the experiments with same parameter setting (*e.g.* layer K=2) for variations of the proposed network in Table 6.

4.5.1 Effects of Region-level Dynamic Graph. In Table 6, we can see that learning by the dynamic graph initialized with prior knowledge (indicated by D_G) outperforms baseline with an improvement of average F1-frame from 63.1% to 65.0%, indicating that the dynamic graph could get richer features from other correlated AU regions to improve robustness. Furthermore, to cancel out the initialization of prior knowledge, we randomly initialize the dynamic graph, which decreases F1-frame to 64.7%. These observations suggest that the relationship reasoning in the dynamic graph can significantly boost the performance of AU detection, while prior knowledge makes a great contribution but not predominantly.

4.5.2 Effects of Multi-level Global Features. We test the contributions of multiple important global feature components of the model in Table 6, namely, original global feature (O_G) from stem network, channel-level global feature (C_G) from channel-level MH-GAT and pixel-level global feature (P_G) from pixel-level MH-GAT. After we supplemented original global feature

Table 7. Performance comparison of MGRR-Net with different iteration step number K on DISFA in terms of F1-frame score (in %).

Layers	AU Index								Avg.
	1	2	4	6	9	12	25	26	
K=1	64.5	58.3	74.9	46.1	**54.4**	75.4	92.3	73.1	67.4
K=2	61.3	62.9	75.8	**48.7**	53.8	**75.5**	**94.3**	**73.1**	**68.2**
K=3	**65.5**	**67.0**	**77.6**	40.0	44.9	75.1	94.0	68.8	66.6

Table 8. Mean error (%) results of different face alignment models on DISFA and BP4D (lower is better).

Datasets	MCL	JAA	JÂANet	**MGRR-Net**
DISFA	7.15	6.30	4.02	**3.95**
BP4D	7.20	6.38	**3.80**	4.01

(O_G) for each target AU, the average F1-frame score has been improved from 65.0% to 65.4%, demonstrating the effectiveness of global detail supplementation. The fusion of channel- and pixel-level global features (C_G and P_G) results in a 0.9% increase, indicating that they make the AU more discriminative than only using the original global features. Comparing the results of the fifth test (with C_G) and the sixth test (with P_G) in Table 6 with the third test, one of the channel-level and pixel-level global features can boost the performance by roughly the same amount. It suggests that by supplementing and training different levels of global features for each AU branch, more global details can be provided to detect AUs in terms of different expressions and individuals.

Finally, the hierarchical gated fusion of multi-level global and local features leads to a significant performance improvement to 68.2% in terms of F1-frame score. It validates that the dynamic relationship of multiple related face regions provides more robustness, while the supplementation of multi-level global features makes the AU more discriminative.

4.5.3 Effects of Layer Number. We evaluate the impact of layer number of our proposed iterative reasoning network. As shown in Table 7, MGRR-Net achieves the averaged F1-frame score of 67.4%, 68.2% and 66.6% on DISFA when the reasoning layer number K is set to 1, 2 and 3 respectively. The averaged F1-frame scores on BP4D dataset are 63.5%, 63.7%, and 63.1% respectively. It achieves the best performance when K=2, and is overfitted when K>2. Finally, the optimal number of layers is 2 for our MGRR-Net on DISFA and BP4D datasets.

4.5.4 Results for Face Alignment. We jointly take face alignment network into our MGRR-Net via auxiliary training, which can provide effective muscle regions corresponding to AUs based on the detected landmarks. Table 8 shows the mean error results of our MGRR-Net and baseline method JÂANet [44] on DISFA and BP4D. We also compare with state-of-the-art face alignment methods that have released trained models, including MCL [46], JAA [43]. Our MGRR-Net achieves competitive 3.95 and 4.01 mean errors on DISFA and BP4D respectively. It indicates that with the comparable face alignment performance as JÂANet, our MGRR-Net can achieve better AU detection accuracy.

4.6 Visualization of Results

To better understand the effectiveness of our proposed model, we visualize the learned class activation maps of MGRR-Net corresponding to different AUs in terms of different expressions, postures and individuals, as shown in Fig. 4. Three examples are from DISFA and three are from BP4D (Two bad

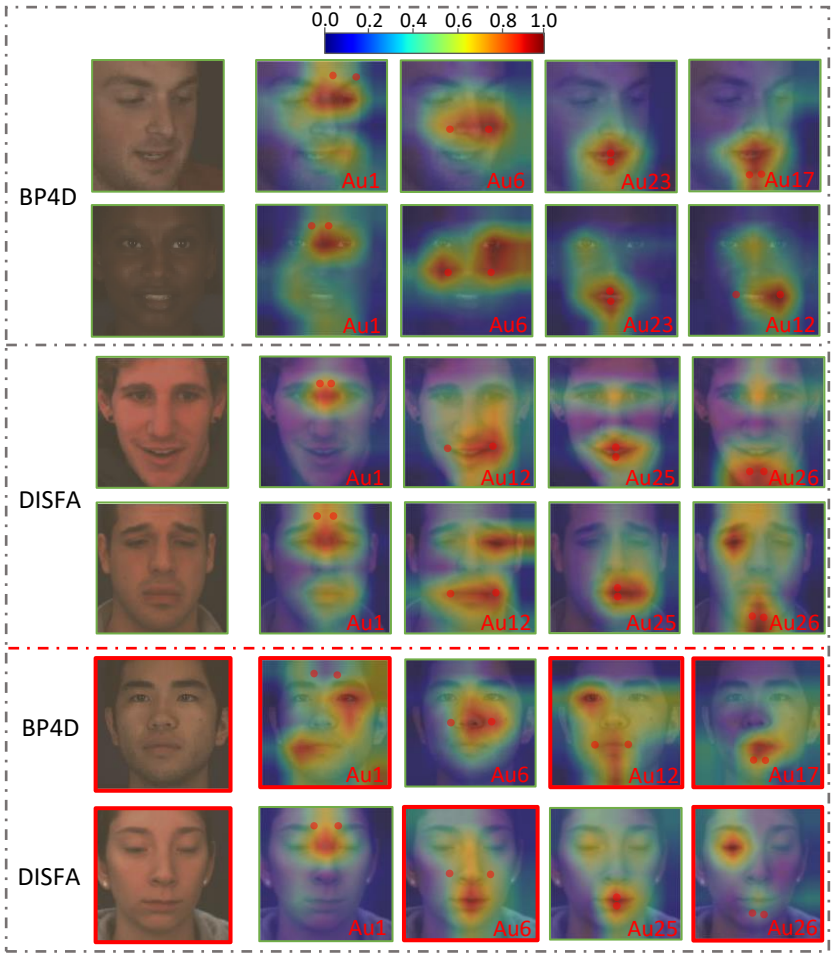

Fig. 4. Class activation maps that show the discriminative regions for different AUs in terms of different expressions and individuals on DISFA and BP4D datasets. We show the region center positions defined by the detected landmarks for the corresponding AUs. Abnormally shifted AU activation maps are marked with red boxes.

examples of abnormal offsets happening are shown at the bottom of Fig. 4.), containing visualization results of different genders and different poses with different AU categories. Through the learning of MGRR-Net, not only the concerned AU regions can be accurately located, but also the positive correlation with other AU areas can be established and other details of the global face can be supplemented. The different activation maps of the same AU on different individuals show that our MGRR-Net can dynamically adjust according to the differences of expression, posture, and individual. Some activation maps are inconsistent with the predefined AU areas, which may be caused by the insensitivity to the target predefined areas after the introduction of multi-level global supplementation. In addition, as shown in Fig. 5, we further visualize the learned relevance matrix (marked as (b)) and the predefined AU correlations (marked as (a)) of the individual corresponding to the first row of Fig. 4 on BP4D. The predefined correlation matrices are used to roughly calculate the co-occurrence relevance between different AUs by counting the dependence of positive and

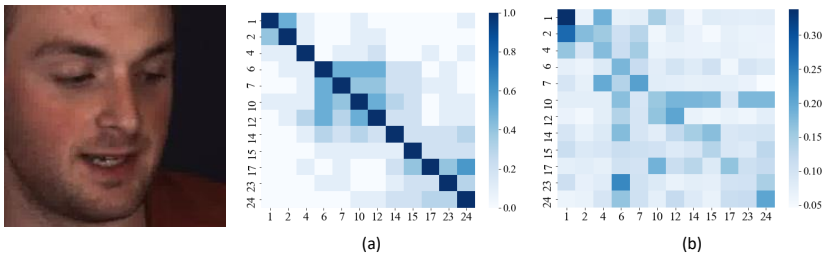

Fig. 5. Visualizations of the predefined AU correlation (a) and the learned relevance matrix (b) for the individual on BP4D. The corresponding class activation maps are shown in the first row of Figure 4.

negative samples. It over-emphasises target AU as well as a few other AUs while other AU regions are completely ignored, due to bias in the statistics of the data. From the correlation matrix we learned, the target AU and the relevant AU are highlighted without discarding information from other branches at all, which is beneficial for increasing the distinguishability between AUs. Furthermore, the supplementation of global features with multiple perspectives allows different AUs to access a lot of information outside the defined areas, as shown in Fig. 4, which is helpful for adaptive changes in terms of different individuals and their expressions.

5 CONCLUSION

In this paper, we have proposed a novel multi-level graph relational reasoning network (termed MGRR-Net) for facial AU detection. Each layer of MGRR-Net can encode the dynamic relationships among AUs via a region-level relationship graph and multiple complementary levels of global information covering expression and subject diversities. The multi-layer iterative feature refinement finally obtains robust and discriminative features for each AU. Extensive experimental evaluations on DISFA and BP4D show that our MGRR-Net outperforms state-of-the-art AU detection methods with impressive margins.

In our future work, we will introduce the pre-trained models to improve the performance of the stem network in extracting feature representation, and we would like to investigate the implementation of facial AU detection into real applications, such as automatically estimating facial palsy severity for patients. This will be helpful for the diagnosis and treatment of people who have facial palsy across the world. In collaboration with medical professionals, we will collect and annotate facial palsy datasets, such as [38], to further validate the migration capability and effectiveness of the proposed model.

ACKNOWLEDGMENTS

This research has been supported in part by the National Natural Science Foundation of China (No. 62176249) and in part by the China Scholarship Council (CSC) from the Ministry of Education of China (No.202006310028).

REFERENCES

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *IEEE FG*. 67–74.

[2] Yingjie Chen, Diqi Chen, Tao Wang, Yizhou Wang, and Yun Liang. 2022. Causal intervention for subject-deconfounded facial action unit recognition. In *AAAI*, Vol. 36. 374–382.

[3] Yuedong Chen, Guoxian Song, Zhiwen Shao, Jianfei Cai, Tat-Jen Cham, and Jianmin Zheng. 2022. GeoConv: Geodesic guided convolution for facial action unit recognition. *Pattern Recognit.* 122 (2022), 108355.

[4] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. 2018. Deep structure inference network for facial action unit recognition. In *ECCV*. 298–313.

[5] Zijun Cui, Chenyi Kuang, Tian Gao, Kartik Talamadupula, and Qiang Ji. 2023. Biomechanics-guided Facial Action Unit Detection through Force Modeling. In *CVPR*. 8694–8703.

[6] Zijun Cui, Tengfei Song, Yuru Wang, and Qiang Ji. 2020. Knowledge augmented deep neural networks for joint facial expression and action unit recognition. *NeurIPS* 33 (2020), 14338–14349.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*. 248–255.

[8] Paul Ekman and Erika L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[9] Xuri Ge, Joemon M Jose, Pengcheng Wang, Arunachalam Iyer, Xiao Liu, and Hu Han. 2023. ALGRNet: Multi-Relational Adaptive Facial Action Unit Modelling for Face Representation and Relevant Recognitions. *IEEE trans. biom. behav. identity sci.* (2023).

[10] Xuri Ge, Pengcheng Wan, Hu Han, Joemon M Jose, Zhilong Ji, Zhongqin Wu, and Xiao Liu. 2021. Local Global Relational Network for Facial Action Units Recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 01–08.

[11] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5-6 (2005), 602–610.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE CVPR*. 770–778.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[14] Geethu Miriam Jacob and Bjorn Stenger. 2021. Facial Action Unit Detection With Transformers. In *IEEE CVPR*. 7680–7689.

[15] Shashank Jaiswal and Michel Valstar. 2016. Deep learning the dynamic appearance and shape of facial action units. In *IEEE WACV*. 1–8.

[16] Xibin Jia, Shaowu Xu, Yuhan Zhou, Luo Wang, and Weiting Li. 2023. A novel dual-channel graph convolutional neural network for facial action unit recognition. *Pattern Recogn. Lett.* 166 (2023), 61–68.

[17] Xibin Jia, Yuhan Zhou, Weiting Li, Jinghua Li, and Baocai Yin. 2022. Data-aware relation learning-based graph convolution neural network for facial action unit recognition. *Pattern Recognit. Lett.* 155 (2022), 100–106.

[18] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[19] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. 2019. Semantic relationships guided representation learning for facial action unit recognition. In *AAAI*. 8594–8601.

[20] Wei Li, Farnaz Abtahi, and Zhigang Zhu. 2017. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *IEEE CVPR*. 1841–1850.

[21] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. 2018. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 11 (2018), 2583–2596.

[22] Xiaobai Li, Jukka Komulainen, Guoying Zhao, Pong-Chi Yuen, and Matti Pietikäinen. 2016. Generalized face anti-spoofing by detecting pulse from face videos. In *IEEE ICPR*. 4244–4249.

[23] Yante Li, Xiaohua Huang, and Guoying Zhao. 2021. Micro-expression action unit detection with spatial and channel attention. *Neurocomputing* 436 (2021), 221–231.

[24] Yong Li and Shiguang Shan. 2023. Contrastive Learning of Person-independent Representations for Facial Action Unit Detection. *IEEE Trans. Image Process.* (2023).

[25] Yongqiang Li, Shangfei Wang, Yongping Zhao, and Qiang Ji. 2013. Simultaneous facial feature tracking and facial expression recognition. *IEEE Trans. Image Process.* 22, 7 (2013), 2559–2573.

[26] Yante Li and Guoying Zhao. 2021. Intra-and Inter-Contrastive Learning for Micro-expression Action Unit Detection. In *ACM ICMI*. 702–706.

[27] Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. 2016. Semantic object parsing with graph lstm. In *ECCV*. Springer, 125–143.

[28] Peng Liu, Zheng Zhang, Huiyuan Yang, and Lijun Yin. 2019. Multi-modality empowered network for facial action unit detection. In *IEEE WACV*. 2175–2184.

[29] Ping Liu, Joey Tianyi Zhou, Ivor Wai-Hung Tsang, Zibo Meng, Shizhong Han, and Yan Tong. 2014. Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis. In *ECCV*. 151–166.

[30] Zhilei Liu, Jiahui Dong, Cuicui Zhang, Longbiao Wang, and Jianwu Dang. 2020. Relation modeling with graph convolutional networks for facial action unit detection. In *MMM*. Springer, 489–501.

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE ICCV*. 10012–10022.

[32] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. 2022. Learning Multi-dimensional Edge Feature-based AU Relation Graph for Facial Action Unit Recognition. In *IJCAI*. 1239–1246.

[33] Chen Ma, Li Chen, and Junhai Yong. 2019. AU R-CNN: Encoding expert prior knowledge into R-CNN for action unit detection. *Neurocomputing* 355 (2019), 35–47.
[34] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. 2013. Disfa: A spontaneous facial action intensity database. *IEEE Trans. Affect. Comput.* 4, 2 (2013), 151–160.
[35] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. 2019. Multi-label co-regularization for semi-supervised facial action unit recognition. In *Ner*. 909–919.
[36] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. 2019. Local relationship learning with person-specific shape regularization for facial action unit detection. In *IEEE CVPR*. 11917–11926.
[37] Xuesong Niu, Hu Han, Jiabei Zeng, Xuran Sun, Shiguang Shan, Yan Huang, Songfan Yang, and Xilin Chen. 2018. Automatic engagement prediction with GAP feature. In *ACM ICMI*. 599–603.
[38] Brian F O'Reilly, John J Soraghan, Stewart McGrenary, and Shu He. 2010. Objective method of assessing and presenting the House-Brackmann and regional grades of facial palsy by production of a facogram. *Otology & Neurotology* 31, 3 (2010), 486–491.
[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Ner*. 8026–8037.
[40] David R Rubinow and Robert M Post. 1992. Impaired recognition of affect in facial expression in depressed patients. *Biological Psychiatry* 31, 9 (1992), 947–953.
[41] Nishant Sankaran, Deen Dayal Mohan, Nagashri N Lakshminarayana, Srirangaraj Setlur, and Venu Govindaraju. 2020. Domain adaptive representation learning for facial action unit recognition. *Pattern Recognit.* 102 (2020), 107127.
[42] Ziqiao Shang, Congju Du, Bingyin Li, Zengqiang Yan, and Li Yu. 2023. MMA-Net: Multi-view Mixed Attention Mechanism for Facial Action Unit Detection. *Pattern Recogn. Lett.* (2023).
[43] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. 2018. Deep adaptive attention for joint facial action unit detection and face alignment. In *ECCV*. 705–720.
[44] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. 2021. JAA-Net: joint facial action unit detection and face alignment via adaptive attention. *IJCV* 129, 2 (2021), 321–340.
[45] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. 2019. Facial action unit detection using attention and relation learning. *IEEE Trans. Affect. Comput.* (2019).
[46] Zhiwen Shao, Hengliang Zhu, Xin Tan, Yangyang Hao, and Lizhuang Ma. 2020. Deep multi-center learning for face alignment. *Neurocomputing* 396 (2020), 477–486.
[47] Zhiwen Shao, Lixin Zou, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. 2020. Spatio-temporal relation and attention learning for facial action unit detection. *arXiv preprint arXiv:2001.01168* (2020).
[48] Jingang Shi, Iman Alikhani, Xiaobai Li, Zitong Yu, Tapio Seppänen, and Guoying Zhao. 2019. Atrial fibrillation detection from face videos by fusing subtle variations. *IEEE Trans. Circuits Syst. Video Technol.* 30, 8 (2019), 2781–2795.
[49] Tengfei Song, Lisha Chen, Wenming Zheng, and Qiang Ji. 2021. Uncertain graph neural networks for facial action unit detection. In *AAAI*. 5993–6001.
[50] Tengfei Song, Zijun Cui, Yuru Wang, Wenming Zheng, and Qiang Ji. 2021. Dynamic Probabilistic Graph Convolution for Facial Action Unit Intensity Estimation. In *IEEE CVPR*. 4845–4854.
[51] Tengfei Song, Zijun Cui, Wenming Zheng, and Qiang Ji. 2021. Hybrid Message Passing With Performance-Driven Structures for Facial Action Unit Detection. In *IEEE CVPR*. 6267–6276.
[52] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. 2018. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comput.* 11, 3 (2018), 532–541.
[53] Sima Taheri, Qiang Qiu, and Rama Chellappa. 2014. Structure-preserving sparse decomposition for facial expression analysis. *IEEE Trans. Image Process.* 23, 8 (2014), 3590–3603.
[54] Gauthier Tallec, Arnaud Dapogny, and Kevin Bailly. 2022. Multi-order networks for action unit detection. *IEEE Trans. Affect. Comput.* (2022).
[55] Yang Tang, Wangding Zeng, Dafei Zhao, and Honggang Zhang. 2021. PIAP-DF: Pixel-Interested and Anti Person-Specific Facial Action Unit Detection Net with Discrete Feedback Learning. In *IEEE ICCV*. 12899–12908.
[56] Yan Tong and Qiang Ji. 2008. Learning bayesian networks with qualitative constraints. In *IEEE CVPR*. 1–8.
[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Ner* 30 (2017).
[58] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR*. 1–12.
[59] Shangfei Wang, Yanan Chang, and Can Wang. 2021. Dual Learning for Joint Facial Landmark Detection and Action Unit Recognition. *IEEE Trans. Affect. Comput.* (2021).
[60] Zhouxia Wang, Tianshui Chen, Jimmy Ren, Weihao Yu, Hui Cheng, and Liang Lin. 2018. Deep reasoning with knowledge graph for social relationship understanding. *arXiv preprint arXiv:1807.00504* (2018).

[61] Xuehan Xiong and Fernando De la Torre. 2013. Supervised descent method and its applications to face alignment. In *IEEE CVPR*. 532–539.
[62] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*.
[63] Huiyuan Yang, Taoyue Wang, and Lijun Yin. 2020. Adaptive Multimodal Fusion for Facial Action Units Recognition. In *ACM MM*. 2982–2990.
[64] Huiyuan Yang, Lijun Yin, Yi Zhou, and Jiuxiang Gu. 2021. Exploiting Semantic Embedding and Visual Feature for Facial Action Unit Detection. In *IEEE CVPR*. 10482–10491.
[65] Mingjing Yu, Huicheng Zheng, Zhifeng Peng, Jiayu Dong, and Heran Du. 2020. Facial expression recognition based on a multi-task global-local network. *Pattern Recognit. Lett.* 131 (2020), 166–171.
[66] Liangfei Zhang, Ognjen Arandjelovic, and Xiaopeng Hong. 2021. Facial Action Unit Detection with Local Key Facial Sub-region Based Multi-label Classification for Micro-expression Analysis. In *ACM MM*. 11–18.
[67] Liangfei Zhang, Xiaopeng Hong, Ognjen Arandjelovic, and Guoying Zhao. 2021. Short and Long Range Relation Based Spatio-Temporal Transformer for Micro-Expression Recognition. *arXiv preprint arXiv:2112.05851* (2021).
[68] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. 2014. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image Vis. Comput.* 32, 10 (2014), 692–706.
[69] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. 2018. Classifier learning with prior probabilities for facial action unit recognition. In *IEEE CVPR*. 5108–5116.
[70] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. 2015. Joint patch and multi-label learning for facial action unit detection. In *IEEE CVPR*. 2207–2216.
[71] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. 2016. Joint patch and multi-label learning for facial action unit and holistic expression recognition. *IEEE Trans. Image Process.* 25, 8 (2016), 3931–3946.
[72] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. 2016. Deep region and multi-label learning for facial action unit detection. In *IEEE CVPR*. 3391–3399.