Learning Multi-dimensional Edge Feature-based AU Relation Graph for Facial Action Unit Recognition

Paper ID 513

Abstract

The activations of Facial Action Units (AUs) are mutually influenced. While the relationship between a pair of AUs can be complex and unique, existing approaches fail to specifically and explicitly represent such cues for each pair of AUs in each face display. This paper proposes an AU relationship modelling approach that deep learns a unique graph to explicitly describe the relationship between each pair of AUs of the target facial display. Our approach first encodes each AU's activation status and its association with other AUs into a node feature. Then, it learns a pair of multi-dimensional edge features to describe multiple task-specific relationship cues between each pair of AUs. During both node and edge feature learning, our approach also considers the influence of the unique facial display on AUs' relationship by taking the full face representation as an input. Experimental results on BP4D and DISFA datasets show that both node and edge feature learning modules provide large performance improvements for CNN and transformer-based backbones, with our best systems achieving the state-of-the-art AU recognition results. Our approach not only has a strong capability in modelling relationship cues for AU recognition but also can be easily incorporated to various backbones. Our **anonymous** PyTorch code is made available.[1]

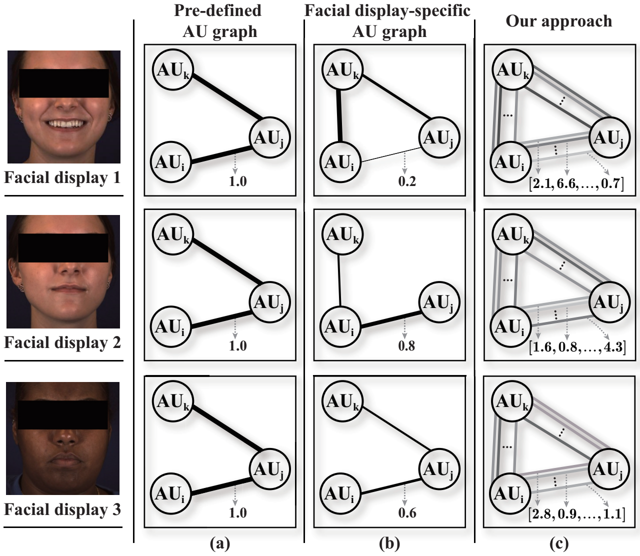

Figure 1: Comparison between our approach with existing AU graph-based approaches: (a) **pre-defined AU graphs** that use a single topology to define AU association for all facial displays; (b) **Facial display-specific AU graphs** that assign a unique topology to define AU association for each facial display. Both (a) and (b) use a single value as an edge feature; (c) **Our approach** encodes a unique AU association pattern for each facial display in node features, and additionally describes the relationship between each pair of AUs using a pair of multi-dimensional edge features.

1 Introduction

Facial Action Coding System [Friesen and Ekman, 1978] represents human face by a set of facial muscle movements called Action Units (AUs). Compared with the emotion-based categorical facial expression model, AUs describe human facial expressions in a more comprehensive and objective manner [Martinez *et al.*, 2017]. Facial AU recognition is a multi-task binary classification problem as multiple AUs can be activated simultaneously. While previous studies found that underlying relationships among AUs' activation [Corneanu *et al.*, 2018; Song *et al.*, 2021b; Shao *et al.*, 2021] are crucial for their recognition, how to properly model such relationships is still an open research question in the field.

A popular strategy applies various traditional machine learning models (*e.g.*, conditional models [Eleftheriadis *et al.*, 2015]) or neural network-based operations (*e.g.*, convolution [Zhao *et al.*, 2016], Long-Short-Term-Memory networks [Niu *et al.*, 2019] or attention [Shao *et al.*, 2021]), to encode all AU descriptors as a single representation which reflects the underlying relationship among all AUs. A key drawback of such solutions is that they fail to individually model the relationship between each pair of AUs, which may contain crucial cues for their recognition (**Problem 1**). Some studies represent all AUs of the target face as a graph, where each AU is represented as a node, and each pair of AUs relation-

[1] https://anonymous.4open.science/r/Anonymous-Github-2DB3

ship is specifically described by an edge that contains a binary value or a single weight to describe their connectivity or association [Song *et al.*, 2021a; Song *et al.*, 2021b]. However, a single value may not be enough to represent the complex underlying relationship between a pair of AUs (**Problem 2**). In particular, some studies [Li *et al.*, 2019; Zhang *et al.*, 2020; Liu *et al.*, 2020] even manually define a single graph topology for all face images based on prior knowledge (*e.g.*, AUs co-occurrence pattern), which fails to consider the influences of the unique facial display on AU relationships (**Problem 3**).

In this paper, we propose a novel AUs relationship modelling approach to address the problems described above, which can be easily incorporated with various deep learning backbones. It takes a full face representation produced by the backbone as the input, and outputs an AUs relation graph that explicitly describes the relationship between each pair of AUs (**addressing problem 1**). Specifically, our approach consists of two modules: (i) the **AUs relationship-aware node feature learning (ANFL) module** first individually learns a representation for each AU from the input full face representation (Sec. 2.1), which encodes not only the AU's activation status but also its association with other AUs; and then (ii) the **multi-dimensional edge feature learning (MEFL) module** learns multiple task-specific relationship cues as the edge representation for each pair of AUs (Sec. 2.2) (**addressing problem 2**). Since both node and edge feature learning take the full face representation as the input, the influence of the unique facial display on AUs relationship is considered when generating its AUs relation graph (**addressing problem 3**).

In summary, the main contributions of our AU relationship modelling approach are that it represents AU relationships as a unique graph for each facial display, which (i) encodes both the activation status of the AU and its association with other AUs into each node feature; (ii) learns a multi-dimensional edge feature to explicitly capture the task-specific relationship cues between each pair of AUs. The main novelty of the proposed approach in comparison to pre-defined AU graphs [Li *et al.*, 2019; Zhang *et al.*, 2020; Liu *et al.*, 2020] and deep learned facial display-specific graphs [Song *et al.*, 2021a; Song *et al.*, 2021b] are illustrated in Figure 1. To the best of our knowledge, this is the first CNN-GCN approach that conducts end-to-end multi-dimensional edge feature learning for face image processing tasks. The pipeline of the proposed approach is illustrated in Figure 2.

2 The proposed approach

Our AU relationship modelling approach deep learns a unique AU relation graph from the target face's the representation, which explicitly captures recognition-related relationship cues among AUs based on the end-to-end learned relationship modelling modules. The learned AU relation graph represents the i_{th} AU as the node $\boldsymbol{v}_i \in \boldsymbol{V}$ in the graph, which contains a vector describing the activation status of the i_{th} AU as well as its association with other AUs in the target facial display. Besides, the task-specific relationship cues between nodes (AUs) $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are also explicitly described by two directed edges $\boldsymbol{e}_{i,j}, \boldsymbol{e}_{j,i} \in \boldsymbol{E}$ that are represented by two deep learned vectors.

2.1 AUs relationship-aware node feature learning

As illustrated in Figure 2, the ANFL module consists of two blocks: an AU-specific Feature Generator (AFG) and a Facial Graph Generator (FGG). The AFG individually generates a representation for each AU, based on which the FGG automatically designs an optimal graph for each facial display, aiming to accurately recognize all target AUs. To achieve this, the FGG would enforce the AFG to encode task-specific associations among AUs into their AU-specific representations.

AU-specific Feature Generator

The AFG is made up of N AU-specific feature extractors, each of which contains a fully connected layer (FC) and a global average pooling (GAP) layer. It takes the full face representation $\boldsymbol{X} \in \mathbb{R}^{C \times D}$ (C channels with D dimensions) as the input, which can be produced by any standard machine learning backbone. The FC layer of i_{th} AU-specific feature extractor first projects the $\boldsymbol{X}$ to an AU-specific feature map $\boldsymbol{U}_i \in \mathbb{R}^{C \times D}$, which is then fed to a GAP layer, yielding a vector containing C values as the i_{th} AU's representation $\boldsymbol{v}_i$. Consequently, N AU representations can be learned from the full face representation $\boldsymbol{X}$, respectively.

Facial Graph Generator

Our hypothesis is that the relationship cues among AUs are unique for each facial display. As a result, directly utilizing relationship cues defined in the training set (*e.g.*, co-occurrence pattern) may not generalise well at the inference stage. As a result, we propose to represent AU relationships in each facial display as a unique graph which considers the influence of the target facial display on AUs relationship.

For a face image, the FGG block treats N target AUs' feature vectors $\boldsymbol{\mathcal{V}} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N\}$ as N node features and defines the connectivity (edge presence) between a pair of nodes $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ by their features' similarity ($\text{Sim}(i,j) = \boldsymbol{v}_i^T \boldsymbol{v}_j$). Specifically, we choose the K nearest neighbours of each node as its neighbours, and thus the graph topology is defined by the learned node features. Then, a GCN layer is employed to jointly update all AUs activation status from the produced graph, where the i_{th} AU's activation representation $\boldsymbol{v}_i^{\text{FGG}}$ is generated by $\boldsymbol{v}_i$ and its connected nodes as:

$$\boldsymbol{v}_i^{\text{FGG}} = \sigma[g(\boldsymbol{v}_i, \sum_{j=1}^{N} r(\boldsymbol{v}_j, a_{i,j}))] \tag{1}$$

where $\sigma[]$ is the non-linear activation; g and r denote differentiable functions of the GCN layer, and $a_{i,j} \in \{0, 1\}$ represents the connectivity between $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$.

To provide a prediction for the i_{th} AU, we propose a similarity calculating (SC) strategy which learns a trainable vector $\boldsymbol{s}_i$ that has the same dimension as the $\boldsymbol{v}_i^{\text{FGG}}$, and then generates the i_{th} AU's occurrence probability by computing the cosine similarity between $\boldsymbol{v}_i^{\text{FGG}}$ and $\boldsymbol{s}_i$ as:

$$p_i^{\text{FGG}} = \frac{\text{ReLU}(\boldsymbol{v}_i^{\text{FGG}})^T \text{ReLU}(\boldsymbol{s}_i)}{\|\text{ReLU}(\boldsymbol{v}_i^{\text{FGG}})\|_2 \|\text{ReLU}(\boldsymbol{s}_i)\|_2} \tag{2}$$

where ReLU denotes a non-linearity activation. As a result, a pair of AUs that have a strong association (high similarity) would have connected nodes. In other words, the FGG block

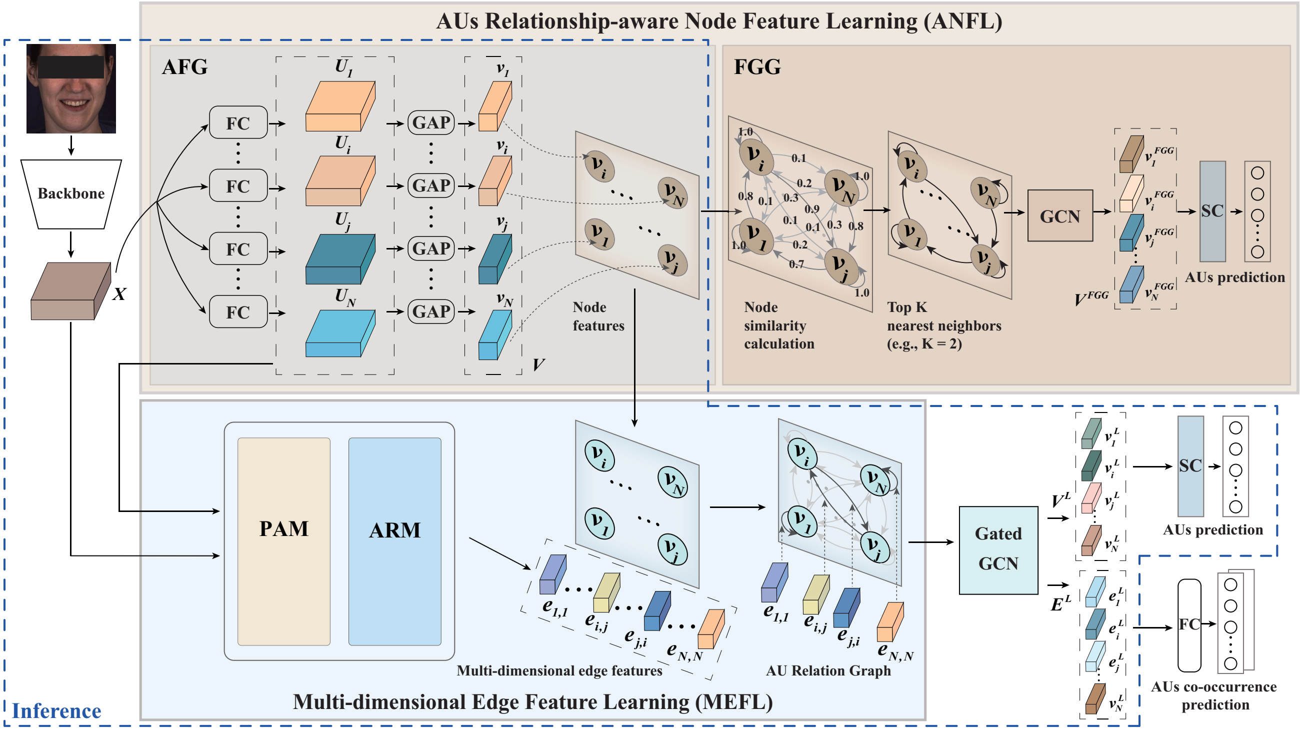

Figure 2: The pipeline of the proposed AU relationship modelling approach. It takes the full face representation X as the input, and the AFG block that is jointly trained with the FGG block, firstly provides a vector as a node feature to describe each AU's activation as well as its association with other AUs (Sec. 2.1). Then, the MEFL module learns a pair of vectors as multi-dimensional edge features to describe task-specific relationship cues between each pair of AUs (Sec. 2.2). The AU relation graph produced by our approach is then fed to a GatedGCN for AU recognition. Only the modules and blocks contained within the blue dashed lines are used at the inference stage.

enforces the AFG block to encode node (AU) features that contain task-specific relationship cues among AUs of the target facial display, in order to produce an optimal graph for their recognition.

2.2 Multi-dimensional edge feature learning

In addition to relationship cues encoded in node features, we also propose a Multi-dimensional Edge Feature Learning (MEFL) module to deep learn a pair of multi-dimensional edge features, aiming to explicitly describe task-specific relationship cues between each pair of AUs. Importantly, we learn edge features for both connected and un-connected node pairs defined in Sec. 2.1. Even when a pair of nodes have low similarity, their relationship may still contain crucial cues for AU recognition, which are ignored during the node feature learning. Since an AU's activation may also influence other AUs' status, the relationship between a pair of AUs can be reflected by not only their features but also AUs defined by other facial regions. Thus, the MEFL module consists of two blocks: a **Facial display-specific AU representation modelling (FAM)** block that first locates activation cues of each AU from the full face representation, and an **AU relationship modelling (ARM)** block that further extracts features from these located cues, which relate to both AUs activation. This is also illustrated in Figure 3.

FAM. As illustrated in Figure 3, for a pair of AUs, the FAM takes their AU-specific feature maps $\boldsymbol{U}_i$, $\boldsymbol{U}_j$, and the full face representation $\boldsymbol{X}$ as the input. It first conducts cross attention between $\boldsymbol{U}_i$ and $\boldsymbol{X}$ as well as $\boldsymbol{U}_j$ and $\boldsymbol{X}$, respectively, where AU-specific feature maps $\boldsymbol{U}_i$ and $\boldsymbol{U}_j$ are individually used as queries, while the full face representation $\boldsymbol{X}$ is treated as the key and value. This process can be formulated as:

$$\boldsymbol{\mathcal{F}}_{i,x}^{AS}, \boldsymbol{\mathcal{F}}_{j,x}^{AS} = \text{FAM}(\boldsymbol{U}_i, \boldsymbol{X}), \text{FAM}(\boldsymbol{U}_j, \boldsymbol{X}) \quad (3)$$

with the cross attention operation in FAM defined as

$$\text{FAM}(\boldsymbol{A}, \boldsymbol{B}) = \text{softmax}(\frac{\boldsymbol{A}\boldsymbol{W}_q(\boldsymbol{B}\boldsymbol{W}_k)^T}{\sqrt{d_k}})\boldsymbol{B}\boldsymbol{W}_v \quad (4)$$

where $\boldsymbol{W}_q$, $\boldsymbol{W}_k$ and $\boldsymbol{W}_v$ are learnable weights, and d_k is a scaling factor equalling to the number of the 'key's' channels. As a result, the produced $\boldsymbol{\mathcal{F}}_{i,x}^{AS}$ and $\boldsymbol{\mathcal{F}}_{j,x}^{AS}$ extract and highlight the most important facial cues from all facial regions of the target facial display for AU i and AU j's recognition, respectively, which consider the influence of the unique facial display on AUs relationships.

ARM. After encoding task-specific facial cues for each AU's recognition independently, the ARM block further extracts the facial cues related to both AUs' recognition. It also conducts the cross-attention (has the same form as Eq. 4 but independent weights) between $\boldsymbol{\mathcal{F}}_{i,x}^{AS}$ and $\boldsymbol{\mathcal{F}}_{j,x}^{AS}$, and produces features $\boldsymbol{\mathcal{F}}_{i,j,x}^{AR}$ and $\boldsymbol{\mathcal{F}}_{j,i,x}^{AR}$, where $\boldsymbol{\mathcal{F}}_{i,j,x}^{AR}$ is generated by using $\boldsymbol{\mathcal{F}}_{j,x}^{AS}$ as the query and $\boldsymbol{\mathcal{F}}_{i,x}^{AS}$ as the key and value, while $\boldsymbol{\mathcal{F}}_{j,i,x}^{AR}$

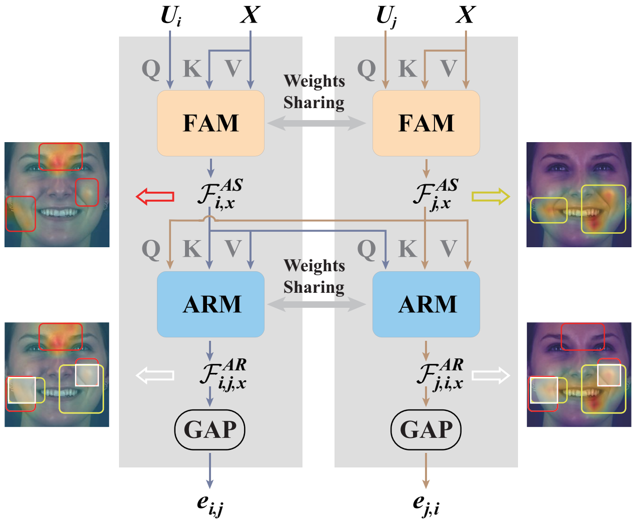

Figure 3: Illustration of the MEFL module. The **FAM** first independently locates activation cues related to i_{th} and j_{th} AU-specific feature maps $\boldsymbol{U}_i$ and $\boldsymbol{U}_j$ in the full face representation $\boldsymbol{X}$ (activated face areas are depicted in red and yellow). Then, the **ARM** further extracts cues related to both $\boldsymbol{U}_i$ and $\boldsymbol{U}_j$ (depicted in white), based on which multi-dimensional edge features $\boldsymbol{e}_{i,j}$ and $\boldsymbol{e}_{j,i}$ are produced.

is generated by using $\mathcal{F}^{AS}_{i,x}$ as the query and $\mathcal{F}^{AS}_{j,x}$ as the key and value. As a result, the $\mathcal{F}^{AR}_{i,j,x}$ summarizes $\mathcal{F}^{AS}_{j,x}$-related cues in the $\mathcal{F}^{AS}_{i,x}$, and $\mathcal{F}^{AR}_{j,i,x}$ summarizes $\mathcal{F}^{AS}_{i,x}$-related cues in the $\mathcal{F}^{AS}_{j,x}$. Finally, we feed $\mathcal{F}^{AR}_{i,j,x}$ and $\mathcal{F}^{AR}_{j,i,x}$ to a GAP layer to obtain multi-dimensional edge feature vectors $\boldsymbol{e}_{i,j}$ and $\boldsymbol{e}_{j,i}$, respectively. Mathematically speaking, this process can be represented as

$$\boldsymbol{e}_{i,j}, \boldsymbol{e}_{j,i} = \text{GAP}(\text{ARM}(\mathcal{F}^{AS}_{j,x}, \mathcal{F}^{AS}_{i,x}), \text{ARM}(\mathcal{F}^{AS}_{i,x}, \mathcal{F}^{AS}_{j,x})) \tag{5}$$

In short, the features encoded in edge features $\boldsymbol{e}_{i,j}$ and $\boldsymbol{e}_{j,i}$ summarize multiple facial cues that relate to both i_{th} and j_{th} AUs' recognition, from all facial regions of the target face.

Once the AUs relation graph $\boldsymbol{G}^0 = (\boldsymbol{V}^0, \boldsymbol{E}^0)$ that consists of N node features and $N \times N$ multi-dimensional directed edge features is learned, we feed it to a GCN model to jointly recognize all target AUs. In this paper, we use a model that only consists of L gated graph convolution layers (GatedGCN) [Bresson and Laurent, 2017], and thus the output $\boldsymbol{G}^L = (\boldsymbol{V}^L, \boldsymbol{E}^L)$ is also a graph that has the same topology as $\boldsymbol{G}^0$, where the i_{th} node $\boldsymbol{v}_i^L$ represents the i_{th} AU's activation status ($L = 2$ in this paper). We finally re-employ the SC module proposed in the FGG block to predict N AUs' activation from the node features of $\boldsymbol{G}^L$. During the inference stage, only the well-trained AFG and MEFL are used to process the input full face representation and generate the AU relation graph.

2.3 Training strategy

In this paper, we propose a two-stage training method to jointly optimize the proposed ANFL and MEFL modules with the backbone and classifier in an end-to-end manner.

In the first stage, we train the backbone with the ANFL module, aiming to learn an AFG block that produces node features containing both AU activation status and their associations for each facial display. To achieve this, we propose a weighted asymmetric loss to compute the loss value between the ground truth and predictions generated by the FGG block. A priori, we notice that existing AU datasets usually have unbalanced labels, where some AUs occurred less frequently than others, and most AUs are inactivated for the majority of face images. To alleviate such issues, our weighted asymmetric loss is formulated as:

$$\mathcal{L}_{\text{WA}} = -\frac{1}{N}\sum_{i=1}^{N} w_i[y_i \log(p_i) + (1-y_i)p_i \log(1-p_i)] \tag{6}$$

where p_i, y_i and w_i are the prediction (occurrence probability), ground truth and weight of the i_{th} AU, respectively. Here, the $w_i = (1/r_i)/\Sigma_{j=1}^{N}(1/r_j)$ is defined by the i_{th} AU's occurrence rate r_i computed from the training set. It allows loss values to account less for AUs that have higher occurrence rates in the training set, leading loss values caused by less frequently occurring AUs to have higher weights during the training. Additionally, the term 'p_i' in the center of $(1-y_i)p_i\log(1-p_i)$ down weights loss values caused by inactivated AUs that are easy to be recognized, whose predicted occurrence probabilities are close to zero ($p_i \ll 0.5$), enforcing the training process to focus on activated AUs and inactivated AUs that are hard to be correctly recognized.

The second stage trains the MEFL module and classifier (GatedGCN) with the pre-trained backbone and AFG block. Here, we again employ the proposed weighted asymmetric loss (Eq. 6) to compute the loss value $\mathcal{L}_{\text{WA}}$ between the outputs of the classifier and ground truth. Additionally, we also leverage the AUs co-occurrence patterns to supervise the training process. We feed multi-dimensional edge features $\boldsymbol{e}^L_{i,j}$ and $\boldsymbol{e}^L_{j,i}$ generated from the last GatedGCN layer to a shared FC layer, in order to predict the co-occurrence pattern of the i_{th} and j_{th} AUs of the target face. We define this task as a four-class classification problem, *i.e.,* for a pair of nodes $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$: (1) both $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are inactivated; (2) $\boldsymbol{v}_i$ is inactivated and $\boldsymbol{v}_j$ is activated; (3) $\boldsymbol{v}_i$ is activated and $\boldsymbol{v}_j$ is inactivated; or (4) both $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are activated. As a result, the categorical cross-entropy loss is introduced as:

$$\mathcal{L}_{\text{E}} = -\frac{1}{|\boldsymbol{E}|}\sum_{i=1}^{|\boldsymbol{E}|}\sum_{j=1}^{N_E} y^e_{i,j}\log\left(\frac{e^{p^e_{i,j}}}{\sum_k e^{p^e_{i,k}}}\right) \tag{7}$$

where $|\boldsymbol{E}|$ denotes the number of edges in the facial graph; N_E is the number of co-occurrence patterns; $p^e_{i,j}$ is the co-occurrence prediction output from the shared FC layer. Consequently, The overall training loss of the second stage is formulated as the weighted combination of the two losses:

$$\mathcal{L} = \mathcal{L}_{\text{WA}} + \lambda\mathcal{L}_{\text{E}} \tag{8}$$

where λ decides the relative importance of the two losses.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate the performance of our approach on two widely-used benchmark datasets: BP4D [Zhang *et al.*,

Method	AU												Avg.
	1	2	4	6	7	10	12	14	15	17	23	24	
EAC-Net [Li *et al.*, 2018]	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
JAA-Net [Shao *et al.*, 2018]	47.2	44.0	54.9	77.5	74.6	84.0	86.9	61.9	43.6	60.3	42.7	41.9	60.0
LP-Net [Niu *et al.*, 2019]	43.4	38.0	54.2	77.1	76.7	83.8	87.2	63.3	45.3	60.5	48.1	54.2	61.0
ARL [Shao *et al.*, 2019]	45.8	39.8	55.1	75.7	77.2	82.3	86.6	58.8	47.6	62.1	47.4	[55.4]	61.1
SEV-Net [Yang *et al.*, 2021]	[**58.2**]	[**50.4**]	58.3	[**81.9**]	73.9	[**87.8**]	87.5	61.6	[52.6]	62.2	44.6	47.6	63.9
FAUDT [Jacob and Stenger, 2021]	51.7	[49.3]	[**61.0**]	77.8	79.5	82.9	86.3	[67.6]	51.9	63.0	43.7	[**56.3**]	64.2
SRERL [Li *et al.*, 2019]	46.9	45.3	55.6	77.1	78.4	83.5	87.6	63.9	52.2	[63.9]	47.1	53.3	62.9
UGN-B [Song *et al.*, 2021a]	[54.2]	46.4	56.8	76.2	76.7	82.4	86.1	64.7	51.2	63.1	48.5	53.6	63.3
HMP-PS [Song *et al.*, 2021b]	53.1	46.1	56.0	76.5	76.9	82.1	86.4	64.8	51.5	63.0	[49.9]	54.5	63.4
Ours (ResNet-50)	53.7	46.9	59.0	78.5	[80.0]	84.4	[87.8]	67.3	52.5	63.2	**50.6**	52.4	[64.7]
Ours (Swin Transformer-Base)	52.7	44.3	[60.9]	[79.9]	[**80.1**]	[85.3]	[**89.2**]	[**69.4**]	[**55.4**]	[**64.4**]	49.8	55.1	[**65.5**]

Table 1: F1 scores (in %) achieved for 12 AUs on BP4D dataset, where the three methods (SRERL, UGN-B and HMP-PS) listed in the middle of the table are also built with graphs. The best, second best, and third best results of each column are indicated with brackets and bold font, brackets alone, and underline, respectively.

Method	AU								Avg.
	1	2	4	6	9	12	25	26	
EAC-Net [Li *et al.*, 2018]	41.5	26.4	66.4	50.7	[**80.5**]	[**89.3**]	88.9	15.6	48.5
JAA-Net [Shao *et al.*, 2018]	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
LP-Net [Niu *et al.*, 2019]	29.9	24.7	72.7	46.8	49.6	72.9	93.8	65.0	56.9
ARL [Shao *et al.*, 2019]	43.9	42.1	63.6	41.8	40.0	76.2	[95.2]	[66.8]	58.7
SEV-Net [Yang *et al.*, 2021]	[**55.3**]	[**53.1**]	61.5	53.6	38.2	71.6	[**95.7**]	41.5	58.8
FAUDT [Jacob and Stenger, 2021]	46.1	[48.6]	72.8	[**56.7**]	50.0	72.1	90.8	55.4	61.5
SRERL [Li *et al.*, 2019]	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
UGN-B [Song *et al.*, 2021a]	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
HMP-PS [Song *et al.*, 2021b]	38.0	45.9	65.2	50.9	50.8	76.0	93.3	[**67.6**]	61.0
Ours (ResNet-50)	[54.6]	47.1	[72.9]	[54.0]	[55.7]	[76.7]	91.1	53.0	[**63.1**]
Ours (Swin Transformer-Base)	52.5	45.7	[**76.1**]	51.8	46.5	76.1	92.9	57.6	[62.4]

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

2014] and DISFA [Mavadati *et al.*, 2013]. BP4D recorded 328 videos (about 140,000 facial frames) from 41 young adults (23 females and 18 males) who were asked to respond to 8 emotion elicitation tasks. DISFA recorded $130,815$ frames from 27 subjects (12 females and 15 males) who were watching Youtube videos. Each frame in BP4D and DISFA is annotated with occurrence labels of multiple AUs.

Implementation Details. For both datasets, we use MTCNN [Yin and Liu, 2017] to perform face detection and alignment for each frame and crop it to 224×224 as the input for backbones. We then follow the same protocol as previous studies [Zhao *et al.*, 2016; Li *et al.*, 2018; Song *et al.*, 2021b] to conduct subject-independent three folds cross-validation for each dataset, and report the average results over 3 folds. During the training, we employ an AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and weight decay of $5e^{-4}$. The number K for choosing nearest neighbors in the FGG is set to 3 and 4 for BP4D and DISFA, respectively. For the hyperparameter λ in Eq. 8, we set it to 0.05 and 0.01 for models based on ResNet and Swin Transformer, respectively. We totally train the proposed model for 40 epochs, including 20 epochs for the first stage (the initial learning rate of $1e^{-4}$) and 20 epochs for the second stage (the initial learning rate of $1e^{-6}$), with a batch size of 64. The cosine decay learning rate scheduler is also used. Both backbones are pretrained on ImageNet [Deng *et al.*, 2009]. All our experiments are conducted using NVIDIA A100 GPUs based on the open-source PyTorch platform.

Evaluation Metric. We follow previous AU occurrence recognition studies using a common metric: frame-based F1 score, to evaluate the performance of our approach, which is denoted as $F1 = 2\frac{P \cdot R}{P+R}$. It takes the recognition precision P and recall rate R into consideration.

3.2 Results and Discussion

Comparison to State-of-the-art Methods. This section compares our best systems of two backbones with several state-of-the-art methods on both datasets. Table 1 reports the occurrence recognition results of 12 AUs on BP4D. It can be observed that the proposed AU relationship modelling approach allows both backbones (ResNet-50 and Swin Transformer-Base (Swin-B)) to achieve superior overall F1 scores than all other listed approaches, with 0.5% and 1.3% average improvements over the state-of-the-art [Jacob and

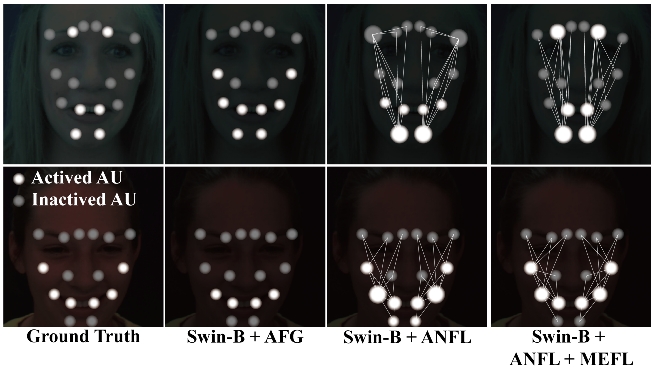

Figure 4: Visualization of association cues encoded in node features (only systems of the last two columns encode such cues). We connect each node to its K nearest neighbours, where nodes of activated AUs usually have more connections than nodes of inactivated AUs. Systems used such relationship cues have enhanced AU recognition results (predictions of the column 3 is better than the column 2).

Stenger, 2021]. Specifically, our approach allows both backbones to achieve the top three performances for 9 out of 12 AUs' recognition (*e.g.*, AU 4, AU 6, AU 7, AU 10, AU 12, AU 14, AU 15, AU 17, and AU 23) among all listed approaches. Similar results were also achieved on DISFA. According to Table 2, our approach helps both backbones to achieve the state-of-the-art average F1 scores over 8 AUs, which outperform the current state-of-the-art with 1.7% and 0.9% improvements, respectively. For fair comparisons, we only compare our approach with static face-based methods that did not remove any frame from the datasets.

According to both tables, our ResNet-50-based system also clearly outperforms other graph-based AU recognition approaches which also use CNNs (ResNet (UGN-B, HMP-PS) or VGG (SRERL)) as backbones. Since SRERL only uses a pre-computed adjacent matrix to describe the relationship between AUs for all faces, our system shows a large advantage over it, with 1.8% and 7.2% F1 score improvements for the average results on BP4D and DISFA, respectively. Although UGN-B and HMP-PS assigned each facial display a unique adjacent matrix and achieved better performance than SRERL, they still use a single value to describe the relationship between each pair of AUs, without considering multiple relationship cues. Thus, our deep-learned task-specific multi-dimensional edge features lead our system to achieve more than 1.3% and 2.1% average F1 score improvements over UGN-B and HMP-PS on both datasets.

Backbone	AFG	FGG	MEFL	$\mathcal{L}_{\mathrm{WA}}$	$\mathcal{L}_{\mathrm{E}}$	**Res**	**Swin**
✓						59.1	62.6
✓	✓					60.4	63.6
✓	✓			✓		63.0	64.6
✓	✓	✓		✓		63.7	65.1
✓	✓		✓	✓		63.9	64.6
✓	✓	✓	✓	✓		64.5	65.4
✓	✓	✓	✓	✓	✓	64.7	65.5

Table 3: Average AU recognition results (F1 scores (in %)) achieved by various settings using two backbones on the BP4D. The systems of the first two rows are trained with widely-used weighted binary cross-entropy loss.

Ablation Studies. Table 3 evaluates the influence of each component of our pipeline on the average AU recognition results. It can be observed that simply using the AFG block to specifically learn a representation for each AU enhanced the performance for both backbones, indicating that the relationship between each AU's activation and the full face representation is unique. Building on the backbone-AFG system, we also found that individually adding the FFG block or MEFL module further increased the recognition performance for both backbones. These results suggest that (i) the FFG block allows the AFG block to encode additional AU recognition-related cues into node features, *i.e.,* we hypothesize that the FFG can help the AFG to learn AUs' relationship cues for their recognition; and (ii) the multi-dimensional edge features learned by the MEFL module provide more task-specific AU relationship cues to improve the recognition performance, which further validates our hypothesis that a single value is not enough to carry all useful relationship cues between a pair of AUs. Figure 4 visualizes the relationship cues encoded in node features. Since muscle movements of an activated AU usually cause subtle changes to facial regions of other AUs, the result (activated AUs usually have much more connections (influence more nodes) than nodes of inactivated AUs) validates that the learned node features contain AUs' association. In short, the proposed approach can provide valuable relationship cues for AU recognition during both node and edge feature learning. More importantly, jointly using FFG and MEFL with our weighted asymmetric loss largely boosted both backbones' recognition capabilities, *i.e.*, 5.6% and 2.9% F1 score improvements over the original backbones, as well as 1.7% and 0.9% improvements over the backbone-AFG systems. Besides the proposed relationship modelling approaches, we show that the two loss functions also positively improved the recognition performance. The weighted asymmetric loss clearly enhanced the performance over the widely-used weighted binary cross-entropy loss, illustrating its superiority in alleviating data imbalance issue. Meanwhile, the proposed AU co-occurrence supervision also slightly enhanced recognition results for both backbones.

4 Conclusion

This paper proposes to deep learn a graph that explicitly represents relationship cues between each pair of AUs for each facial display. These relationship cues are encoded in both node features and multi-dimensional edge features of the graph. The results demonstrate that the proposed node and edge feature learning methods extracted reliable task-specific relationship cues for AU recognition, *i.e.*, both CNN and transformer-based backbones have been largely enhanced, and achieved state-of-the-art results on two widely used datasets. Since our graph-based relationship modelling approach can be easily incorporated with standard CNN/transformer backbones, it can be directly applied to enhance the performance of multi-label tasks or tasks whose data contains multiple objects, by explicitly exploring the task-specific relationship cues among labels or objects.

References

[Bresson and Laurent, 2017] Xavier Bresson and Thomas Laurent. Residual gated graph convnets. *arXiv preprint arXiv:1711.07553*, 2017.

[Corneanu *et al.*, 2018] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. Deep structure inference network for facial action unit recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 298–313, 2018.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Eleftheriadis *et al.*, 2015] Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *Proceedings of the IEEE international conference on computer vision*, pages 3792–3800, 2015.

[Friesen and Ekman, 1978] E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2):5, 1978.

[Jacob and Stenger, 2021] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021.

[Li *et al.*, 2018] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2583–2596, 2018.

[Li *et al.*, 2019] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8594–8601, 2019.

[Liu *et al.*, 2020] Zhilei Liu, Jiahui Dong, Cuicui Zhang, Longbiao Wang, and Jianwu Dang. Relation modeling with graph convolutional networks for facial action unit detection. In *International Conference on Multimedia Modeling*, pages 489–501. Springer, 2020.

[Martinez *et al.*, 2017] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, 10(3):325–347, 2017.

[Mavadati *et al.*, 2013] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[Niu *et al.*, 2019] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11917–11926, 2019.

[Shao *et al.*, 2018] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European conference on computer vision (ECCV)*, pages 705–720, 2018.

[Shao *et al.*, 2019] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Facial action unit detection using attention and relation learning. *IEEE transactions on affective computing*, 2019.

[Shao *et al.*, 2021] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaa-net: joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129(2):321–340, 2021.

[Song *et al.*, 2021a] Tengfei Song, Lisha Chen, Wenming Zheng, and Qiang Ji. Uncertain graph neural networks for facial action unit detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 1, 2021.

[Song *et al.*, 2021b] Tengfei Song, Zijun Cui, Wenming Zheng, and Qiang Ji. Hybrid message passing with performance-driven structures for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6267–6276, June 2021.

[Yang *et al.*, 2021] Huiyuan Yang, Lijun Yin, Yi Zhou, and Jiuxiang Gu. Exploiting semantic embedding and visual feature for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10482–10491, 2021.

[Yin and Liu, 2017] Xi Yin and Xiaoming Liu. Multitask convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing*, 27(2):964–975, 2017.

[Zhang *et al.*, 2014] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[Zhang *et al.*, 2020] Zheng Zhang, Taoyue Wang, and Lijun Yin. Region of interest based graph convolution: A heatmap regression approach for action unit detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2890–2898, 2020.

[Zhao *et al.*, 2016] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.