Multi-scale Dynamic and Hierarchical Relationship Modeling for Facial Action Units Recognition

Zihan Wang[1,2,3], Siyang Song[4*], Cheng Luo[5], Songhe Deng[1,2,3], Weicheng Xie[1,2,3] and Linlin Shen[1,2,3*]
[1]Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University,
[2]Shenzhen Institute of Artificial Intelligence and Robotics for Society,
[3]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University,
[4]University of Leicester, [5]Monash University

Abstract

Human facial action units (AUs) are mutually related in a hierarchical manner, as not only they are associated with each other in both spatial and temporal domains but also AUs located in the same/close facial regions show stronger relationships than those of different facial regions. While none of existing approach thoroughly model such hierarchical inter-dependencies among AUs, this paper proposes to comprehensively model multi-scale AU-related dynamic and hierarchical spatio-temporal relationship among AUs for their occurrences recognition. Specifically, we first propose a novel multi-scale temporal differencing network with an adaptive weighting block to explicitly capture facial dynamics across frames at different spatial scales, which specifically considers the heterogeneity of range and magnitude in different AUs' activation. Then, a two-stage strategy is introduced to hierarchically model the relationship among AUs based on their spatial distribution (i.e., local and cross-region AU relationship modelling). Experimental results achieved on BP4D and DISFA show that our approach is the new state-of-the-art in the field of AU occurrence recognition. Our code is publicly available at https://github.com/CVI-SZU/MDHR.

1. Introduction

Facial Action Coding System (FACS) [11] specifies a set of Facial Action Units (AUs) to describe multiple atomic human facial muscle movements, which can comprehensively and objectively describe various human facial expressions in an anonymous and concise manner [30]. Recent studies frequently show that AUs are robust and effective low-dimensional facial descriptors for various human behaviours understanding tasks, such as emotion [33, 51], mental health [34, 44] and pain level [10] analysis. As a result, a large number of studies attempt to automatically recognize AU occurrences from facial images or videos [7, 8, 16, 41, 45, 46, 55].

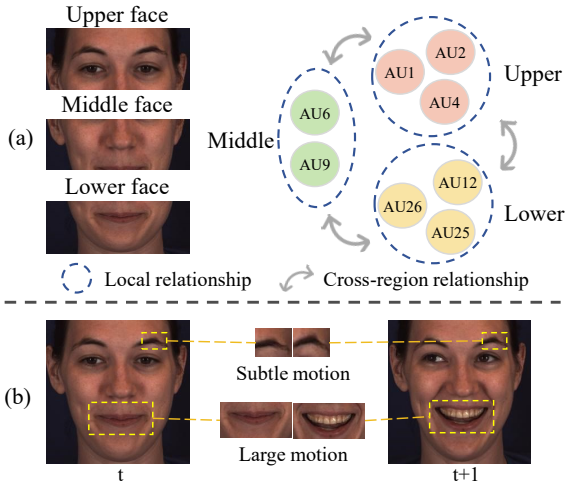

Figure 1. (a) hierarchical AU relationship; and (b) heterogeneous range and magnitude of different AUs' activation.

Most of these approaches conduct AU recognition on still face images. Since each AU usually occur in a specific facial region, some recognize each AU based on a small facial patch defined by automatically detected facial landmarks [14, 18, 38]. However, they not only ignore contextual cues (i.e., AUs are mutually dependent [48]) obtainable from other facial regions for each AU's recognition, but also suffer from errors caused by facial landmark detection. Consequently, other approaches jointly recognize multiple AUs from the entire face, allowing informative contextual cues [32, 36] to be utilized at the cost of including noises introduced by irrelevant facial regions when recognizing a particular AU. Specifically, transformer [13, 57] and graph-based [29, 46] approaches have been widely extended to model the relationship among AUs. However, most of these employ the same strategy to model the relationship between

* Corresponding author

every pair of AUs in the spatial domain(e.g., via transformer encoder with the self-attention operation [13] and graph edges learned by the same cross-attention operation [29]), without giving explicit consideration to the natural hierarchical relationship among AUs (**Problem 1**). More specifically, AUs corresponding to the same/close facial regions frequently show stronger associations than AUs located in different facial regions(illustrated in Fig. 1 (a)), as AUs localized to the same/close facial regions may be influenced by some shared facial muscles [1].

Besides the spatial cues, some studies additionally model temporal dynamic between facial frames to enhance AU recognition performances [23, 24, 47]. A typical solution is applying common temporal models (e.g., Long-Short-Term-Memory (LSTM) [3, 18], Spatio-Temporal Graph [37, 50] and 3D Convolution Neural Networks (CNNs) [5]) to process the extracted frame-level static facial/AU features. However, these temporal modeling strategies are insensitive to subtle facial muscle movements. While other approaches [12, 43, 56] (e.g., optical flow and dynamic image) can explicitly capture facial motions, they still fail to consider that facial muscle movements corresponding to different AUs' activation could exhibit heterogeneity in both range and magnitude (**Problem 2**), e.g., AU25 involves large-scale deformations of the mouth region, while AU2 are represented by subtle muscle movements surrounding eyebrows (illustrated in Fig. 1 (b)). In other words, facial dynamic of a certain spatial scale could contribute unequally to the recognition of different AUs.

In this paper, we propose a novel Multi-scale Dynamic and Hierarchical Relationship (MDHR) modeling approach for AU recognition, which: (i) hierarchically models spatio-temporal relationship among AUs; and (ii) adaptively considers facial dynamic at various spatial scales for each AU's recognition. Our MDHR consists of two key modules. The **Multi-scale Facial Dynamic Modelling (MFD)** module that adaptively emphasizes AU-related facial dynamic at multiple spatial scales (i.e., computing differences between neighboring frames' features maps output from different backbone layers), ensuring both obvious and subtle AU-related facial dynamic can be captured in an efficient manner (**addressing Problem 2**). Then, a **Hierarchical Spatio-temporal AU Relationship Modelling (HSR)** module is introduced to hierarchically model relationship among spatio-temporal AU features in a two-stage manner, where the first stage individually models relationship among AUs within the same/close facial region at both feature extraction and AU prediction levels, and the second stage explicitly learns the relationship between pairs of AUs located in different facial regions via graph edges (**addressing Problem 1**). The main contributions and novelties of this paper are summarised as follows:

- The proposed MFD is the first module that adaptively/specifically considers facial dynamic corresponding to each AU at each spatial scale, as each AUs' activation exhibit heterogeneity in both range and magnitude.
- The proposed HSR is the first module that hierarchically learns local and cross-regional spatio-temporal relationship, while previous approaches fail to consider such hierarchical relationship.
- Experimental results show that our MDHR is the new state-of-the-art on the widely-used AU recognition benchmark datasets: BP4D [59] and DISFA [31], where the proposed MFD and HSR modules positively and complementarily contributed to this decent performance.

2. Related Work

Static face image-based methods: Existing approaches frequently predict AUs' status based on static facial displays. Given the anatomical definition of AUs, many of them [6, 13, 14, 18, 19, 42] attempted to recognize each AU based on a face patch defined by automatically detected facial landmarks or other prior settings. For example, Zhao et al. [61] proposed a patch-based DRML that learns AU representations robust to variations inherent within local facial regions. EAC-Net [19] proposed a cropping layer to learn individual AU's representation from small AU-specific areas. Furthermore, JAA-Net [38] jointly conducted AU recognition and face alignment, where the predicted facial landmarks are used to localize each AU region. To take global facial contextual cues into consideration, alternative approaches [23, 29, 36, 39, 40] learn each AU's representation holistically from the full face image, where spatial attention mechanisms have been widely explored. Shao et al. [36] employed adaptive channel-wise and spatial attention strategy to enforce the model focusing on AU-related local features from the global face. Li et al. [20] proposed a self-diversified multi-channel attention to seek a more robust attention between the global facial representation and each target AU. As AUs are mutually related [60], recent approaches [2, 13, 32, 57] also specifically modelled the underlying relationship among them. For exampple, LP-Net [32] applied LSTMs to capture AU relationship. Jacob et al.[13] proposed a transformer-style AU correlation network. In addition, graph-based strategies have been frequently investigated to model AU relationship[17, 26, 29, 46], where graph nodes have been frequently used to represent target AUs while edges explicitly define the relationship between every pair of AUs.

Spatial-temporal methods: Since facial dynamic also provide crucial cues for AU recognition [4], LSTM has been frequently employed by early studies [4, 15] to model temporal dynamic between static facial features extracted from adjacent frames. To further explore spatio-temporal relationship among AUs, a Spatio-temporal Graph Neural Network (GNN) [37] and a Heterogeneous Spatio-temporal Re-

lation Learning Network (HSTR-Net) [47] have been proposed, both of which first construct a set of spatial graphs to model static AU relationship at the frame-level, and then individually model each AU's temporal dynamic by considering its corresponding spatial graph node features across all frames. In addition, Li et al. [21] applied a transformer to learn both spatial AU dependencies and temporal inter-frame contexts by representing the inter-AU and inter-frame correlations within a multi-head attention matrix. Besides such standard temporal model-based feature-level dynamic modelling, other solutions [12, 22, 43, 53, 54] also have been investigated. For example, two auxiliary AU related tasks (e.g., ROI inpainting and optical flow estimation) are jointly conducted in [52] to enhance the regional features and encode the facial dynamic into the global facial representation, respectively. More recently, Yang et al. [56] introduced a temporal difference network (TDN) that extract facial dynamic at a specific spatial scale. Despite the progress made by approaches discussed above, to the best of our knowledge, none of them has specifically modelled AU-related multi-scale facial dynamic and the hierarchical spatio-temporal relationship among AUs.

3. Methodology

Overview: Given T consecutive facial frames $S = \{f^1, f^2, \cdots, f^T\} \in \mathbb{R}^{T\times C\times H\times W}$, our approach jointly predicts multiple (N) AUs' occurrence at the t_{th} facial frame f^t ($t = 1, 2, \cdots, T$) by taking not only the f^t but also its adjacent frames into consideration. As illustrated in Fig. 2 and Algorithm 1, our MDHR starts with utilizing a backbone (e.g., CNN or Transformer) to jointly extract static facial features from f_t and its adjacent frames $A^t = \{f^{t-k}, \cdots, f^{t-1}, f^{t+1}, \cdots, f^{t+k}\}$. For each frame, multi-scale static facial features are produced by $L-1$ backbone hidden layers and the output layer (the L_{th} layer). Thus, L static facial feature sets corresponding to $2k+1$ frames $X_l = \{x_l^{t-k}, \cdots, x_l^t, \cdots, x_l^{t+k} | l = 1, 2, \cdots, L\}$ are generated (**line 2 in Algorithm 1**). Then, these multi-scale features are fed to the **Multi-scale Facial Dynamic Modelling (MFD)** module, targeting at not only explicitly capturing facial dynamic at multiple spatial scales, but also adaptively combining these multi-scale facial dynamic features with the static feature x_L^t (**line 3 in Algorithm 1**). Based on the spatio-temporal full face representation G^t learned by MFD, a **Hierarchical Spatio-temporal AU Relationship Modelling (HSR)** module further adaptively models the hierarchical spatio-temporal relationship among AUs in a two-stage manner, where the spatial distribution of the target AUs on the human face is considered, resulting in N individual AU representations $\hat{V}^t = \{\hat{v}_1^t, \cdots, \hat{v}_N^t\}$ (**line 4 in Algorithm 1**). Finally, a Temporal Convolution Networks (TCN) [27] with similarity calculating (SC) strategy [29] are employed to predict N AUs' occurrences of the input T frames as $P^1, P^2, \cdots, P^T$ ($P^t = \{p_1^t, \cdots, p_N^t\}$, **line 6 in Algorithm 1**).

Algorithm 1 Pipeline of the proposed approach (MDHR)

Input : T consecutive facial frames $S = \{f^1, \cdots, f^T\}$
Output: N AU's predictions of each frame f^t

1: **for** $t = 1$ **to** T **do**
2: Generating multi-scale static global representations $X_1, X_2 \cdots X_L \leftarrow \text{Backbone}(f^{t-k}, \cdots, f^t \cdots f^{t+k})$
3: Generating global spatio-temporal features $G^t \leftarrow \text{MFD}(X_1, X_2 \cdots X_L)$
4: Generating hierarchical spatio-temporal relationship-aware AU features $\hat{V}^t \leftarrow \text{HSR}(G^t)$
5: **end for**
6: Predicting N AUs of all frames $P^1, P^2, \cdots, P^T \leftarrow \text{SC}(\text{TCN}(\hat{V}^1, \cdots, \hat{V}^t, \cdots, \hat{V}^T))$

3.1. Multi-scale facial dynamic modelling

Inspired by the fact that facial muscle movements are continuous and smooth while each AU exhibit heterogeneity in their range of motions and magnitudes [1], we propose a novel MFD module to model the preceding and proceeding temporal evolution of the target face at multiple spatial scales. It includes a multi-scale Temporal Differecing block that first computes differences between global facial features extracted from every pair of neighboring frames at multiple spatial scales. The obtained multi-scale facial dynamic features are then masked by a set of weighting matrices learned by our adaptive weighting block, aiming to emphasize the informative cues for target AUs at multiple spatio-temporal scales.

Multi-scale Temporal Differecing block: This block is made up of multiple Temporal Differecing (TD) layers followed by convolution layers, which takes feature maps $X_l = \{x_l^{t-k}, \cdots, x_l^t, \cdots, x_l^{t+k} | l = 1, 2, \cdots, L\}$ produced by multiple ($L-1$) hidden layers and the output layer of the backbone as the input, where $x_l^t \in \mathbb{R}^{C_l\times H_l\times W_l}$ denotes the feature map corresponding to the t_{th} facial frame generated from the l_{th} backbone hidden layer (i.e., C_l, H_l, and W_l represent the channel, height and width of the x_l^t, respectively). Here, the l_{th} TD layer conducts point-to-point subtraction on feature maps produced by the l_{th} hidden layer between neighboring frames, aiming to capture facial dynamic at a certain spatial scale. This can be formulated as:

$$d_l^t = x_l^t - x_l^{t-1} \tag{1}$$

Thus, a dynamic feature map $d_l^t \in \mathbb{R}^{C_l\times H_l\times W_l}$ representing the facial dynamic between f^t and f^{t-1} at the l_{th} spatial scale are produced from the l_{th} TD layer. As a result, L sets of dynamic features $D_l = \{d_l^{t-k+1}, \cdots, d_l^t, \cdots, d_l^{t+k} | l = 1, 2, \cdots, L\}$ are obtained to represent facial dynamic at L different scales. After that, we introduce L step convolution

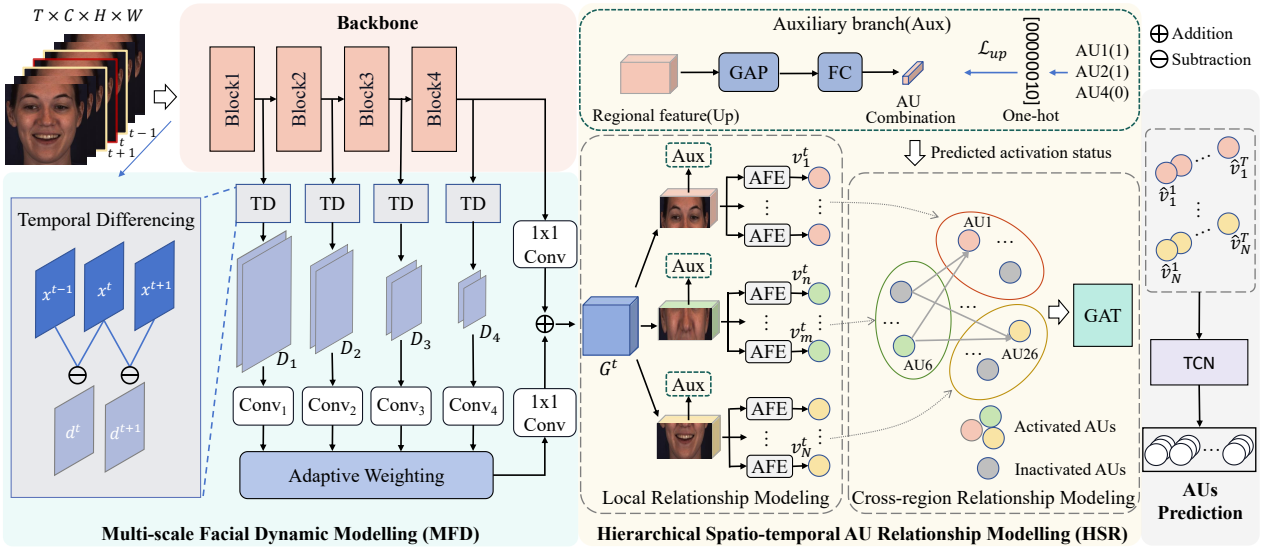

Figure 2. The pipeline of our MDHR, where k is set to 1. The MFD module (Sec. 3.1) first computes facial dynamic at multiple spatial scales based on feature maps output from multiple backbone hidden layers and the output layer. Then, the HSR module (Sec. 3.2) then individually models the relationship among AUs located in the same and different facial regions (the Auxiliary branch is only used at the training phase to make AU combination for each facial region (upper facial region is used as an example in the figure)). Finally, a TCN is individually employed to process every AU feature's sequence of all the input T frames.

layers to resize the dynamic features extracted at different spatial scales as:

$$\hat{d}_l^t = \text{Conv2D}_l(d_l^t) \tag{2}$$

where the kernel size and stride of the l_{th} Conv2D layer Conv2D_l are set to $8/l$, ensuring all produced dynamic features $\hat{d}_l^t \in \mathbb{R}^{c,h,w}$ to have the same shape. Finally, an average pooling is employed to process all re-shaped dynamic features at each spatial scale along the temporal axis as:

$$\bar{d}_l^t = \text{Avg}(\hat{d}_l^{t-k+1}, \cdots, \hat{d}_l^t, \cdots, \hat{d}_l^{t+k}) \tag{3}$$

This way, multi-scale and equal-shape facial dynamic features $\bar{d}_1^t, \bar{d}_2^t, \cdots, \bar{d}_L^t$ of the target frame f^t can be obtained, where each $\bar{d}_l^t$ summarizes the temporal evolution of the f^t by considering its preceding and succeeding k frames.

Adaptive weighting block: Facial muscle movements of large range and magnitude are typically associated with feature maps produced from deep backbone layers while subtle facial dynamic usually can be better described by feature maps produced from shallow backbone layers [25]. Thus, instead of simply conducting element-wise summation or concatenation (i.e., equally treats all components of all feature maps), we propose to adaptively learn L weighting matrices for properly combining the obtained L-scale dynamic features according to the target AUs' typical and unique spatio-temporal scales. In particular, the weight matrix w_l^t at each spatial scale is obtained by exploring the underlying and internal cues from the obtained multi-scale dynamic features, which can be formulated as:

$$w_l^t = \text{Softmax}(\text{Conv}_l(\text{Concat}([\bar{d}_1^t, \bar{d}_2^t, \cdots, \bar{d}_L^t]))) \tag{4}$$

where $l = 1, 2, \cdots, L$. Specifically, multi-scale spatio-temporal features $\bar{d}_1^t, \cdots, \bar{d}_L^t$ of the f_t are first concatenated along their channels, followed by 1×1 convolutions to reduce the number of its channels to one. This results in a unique weighting matrix $w_l^t \in \mathbb{R}^{h \times w}$ to mask the spatio-temporal feature at each spatial scale l. A Softmax function is also applied to normalize the obtained weights such that $\sum_{l=1}^{L} w_l^{t,i,j} = 1$ and $w_l^{t,i,j} \in [0, 1]$, where i and j index the spatial dimensions. Consequently, each obtained weight matrix w_l^t is applied to the corresponding dynamic feature map $\bar{d}_l^t$ by performing element-wise multiplication as:

$$x_{\text{motion}}^t = \sum_{l=1}^{L} w_l^t * \bar{d}_l^t \tag{5}$$

where x_{motion}^t represents the aggregated and adaptively weighted multi-scale facial dynamic representation of the f_t, which is then combined with the spatial feature x_L^t produced by the output layer via the element-wise summation:

$$G^t = x_{\text{motion}}^t + x_L^t \tag{6}$$

In summary, the proposed MFD module adaptively incorporates AU-aware facial dynamic with static and global facial cues into G_t for the fine-grained facial AU recognition.

3.2. Hierarchical spatio-temporal AU relationship modelling

Our HSR module hierarchically models the spatio-temporal relationship among AUs by specifically considering their spatial distribution on the face, as association among AUs in the same/close facial region could be stronger than AUs located in different facial regions [1]. It consists of two stages: the **local AU relationship modelling** stage first models the relationship among AUs located in the same facial region, and then the **cross-regional AU relationship modeling** stage adaptively explore the relationship between AU pairs of different facial regions.

Local AU Relationship Modelling: This stage specifically models relationship among AUs located in the same facial regions at both their features extraction level and prediction level. It builds on the assumption that constraining each AU feature's extraction to its spatially correlated facial regions could partially avoid the negative impacts/noises caused by irrelevant facial regions [19]. Particularly, it first divides the spatio-temporal facial feature $G^t \in \mathbb{R}^{c\times h\times w}$ extracted by MFD module into three subsets corresponding to three slightly overlapped facial regions: (1) the upper region encompassing eyebrows and eyes; (2) the middle region containing the nose and cheeks; and (3) the lower region covering the mouth and chin (Illustrated in Fig. 1). This is achieved by directly slicing the feature G^t along the height dimension as:

$$G^t_{\text{up}}, G^t_{\text{mid}}, G^t_{\text{low}} = G^t[0:\frac{3}{7}h], G^t[\frac{2}{7}h:\frac{5}{7}h], G^t[\frac{4}{7}h:h] \tag{7}$$

where the height h of the G^t is 7 in our implementation, thus we empirically choose this best partition setting. After that, N AU-specific Feature Extractors (**AFE**) (each is made up of a convolution layer with kernel size of 1×1 and a Global Average Pooling (GAP) layer) are employed, where the n_{th} extractor learns a local relationship-aware feature $v^t_n \in \mathbb{R}^{1\times b}$ (b denotes the dimension of an AU vector) from its corresponding sliced regional feature ($G^t_{\text{up}}, G^t_{\text{mid}}$ or G^t_{low}), representing the n_{th} AU's status at the t_{th} frame. Consequently, each AU feature is extracted in the context of its spatially adjacent AUs (i.e., modelling AU relationship of the same facial region at the feature extraction level).

In addition, the spatio-temporal relationship among AUs of the same facial region are also modelled at their prediction level, where an auxiliary branch (**Aux**) is added at the training phase. It is trained to predict an AU occurrence combination $Y^t_{\text{sub}} = \{y^t_{\text{sub},1}, \cdots, y^t_{\text{sub},2^{N_{\text{sub}}}}\}$ (i.e., Y^t_{sub} is a one-hot vector and N_{sub} is the number of the target AUs in the corresponding sub-region) from each sliced regional feature $G^t_{\text{sub}} \in \{G^t_{\text{up}}, G^t_{\text{mid}}, G^t_{\text{low}}\}$, which jointly describes all AUs' occurrence status within each facial region. Mathematically, this process can be formulated as:

$$P^t_{\text{sub}} = \sigma(\text{FC}_{\text{sub}}(\text{GAP}(G^t_{\text{sub}}))) \tag{8}$$

where σ denotes the Softmax function and FC_{sub} denotes a fully connected layer. As a result, training this branch enforces the network encoding underlying local AU relationship to each sliced regional feature, allowing AFE to extract enhanced AU-relevant features from regional features.

Cross-regional AU relationship modeling: Besides spatially adjacent AUs, each AU's activation may also associate with AUs located in other facial regions [17]. Consequently, this stage aims to enhance the recognition performance by additionally capturing such cross-regional AU spatio-temporal dependencies within the given face image. It treats each local relationship-aware spatio-temporal AU feature v^t_n extracted in the previous stage as a node, and adaptively connects it with all activated AU nodes belonging to other facial regions (i.e., AU activation status are decided by AU predictions of the first stage). This edge connection definition is inspired by the finding that activated AUs usually have more influences on other AUs [29]. As a result, the relationship of each cross regional AU pair is explicitly represented through a graph edge, and further modelled via a Graph Attention Network (GAT) [49] layer as:

$$\begin{aligned} e^t_{n,m} &= \text{LeakyReLU}\left(r^T\left[Wv^t_n \parallel Wv^t_m\right]\right) \\ \hat{v}^t_n &= \phi\left(\sum_{m\in N^t_n} \alpha^t_{n,m} W v^t_m\right) \\ \textbf{Subject to: } \alpha^t_{n,m} &= \frac{\exp\left(e^t_{n,m}\right)}{\sum_{q\in N^t_n}\exp\left(e^t_{n,q}\right)} \end{aligned} \tag{9}$$

where $e^t_{n,m}$ is a graph edge defines the impacts of the m_{th} AU node to the n_{th} AU node in the t_{th} frame; $W \in \mathbb{R}^{b\times b}$ denotes a shared linear transformation applied to every node feature; $\parallel$ is the concatenation operation; ϕ is an activation function and $r \in \mathbb{R}^{2b}$ denotes the weight of an attention operation.N^t_n is the set of the neighbours of the current node. Subsequently, N local and global hierarchical relationship-aware AU features $\hat{V}^t = \{\hat{v}^t_1, \hat{v}^t_2, \cdots, \hat{v}^t_N\}$ are generated to describe N target AUs in the t_{th} frame.

3.3. Loss function

As AU recognition constitutes a multi-label binary classification task, with most AUs inactivated the across majority of frames (please refer to Supplementary Material for details), an asymmetric loss function [29] is employed. Given the input consecutive T facial frames with N target AUs, the loss function $\mathcal{L}_{\text{AU}}$ for supervising all AUs' recognition (i.e., output by the TCN/SC layers) is defined as:

$$\mathcal{L}_{\text{AU}} = -\sum_{n=1}^{N}\sum_{t=1}^{T} w_n[y^t_n \log(p^t_n) + p^t_n(1-y^t_n)\log(1-p^t_n)] \tag{10}$$

Method		AU												Avg.
		1	2	4	6	7	10	12	14	15	17	23	24	
Static image-based	EAC-Net [19]	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 [35]	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
	ARL [36]	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
	SMA-Net [20]	56.5	45.1	57.0	79.5	79.5	84.5	86.4	66.1	55.8	64.2	48.7	56.8	64.9
Static AU relationship modeling	SRERL [17]	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
	FAUDT [13]	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
	FAN-Trans [57]	55.4	46.0	59.8	78.7	77.7	82.7	88.6	64.7	51.4	**65.7**	50.9	56.0	64.8
	ME-GraphAU [29]	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
Spatio-temporal	STRAL [37]	48.2	47.7	58.1	75.8	78.1	81.6	87.6	60.5	50.2	64.0	51.2	55.2	63.2
	HSTR-Net[47]	55.5	49.5	**61.9**	76.6	80.2	84.2	87.4	62.6	54.8	64.1	47.1	52.1	64.7
	KS [21]	55.3	48.6	57.1	77.5	**81.8**	83.3	86.4	62.8	52.3	61.3	**51.6**	58.3	64.7
	WSRTL [52]	**59.7**	**51.7**	61.6	**80.3**	80.9	85.2	**89.7**	67.8	52.2	63.4	51.4	46.9	65.9
	Ours (ResNet-50)	58.3	50.9	58.9	78.4	80.3	84.9	88.2	**69.5**	56.0	65.5	49.5	**59.3**	**66.6**
	Ours (Swin-B)	54.6	49.7	61.0	79.9	79.4	**85.4**	88.5	67.8	**56.8**	63.2	50.9	55.4	66.1

Table 1. F1 scores (in %) achieved for 12 AUs on BP4D dataset. The best and the second best results of each column are indicated with bold font and underline, respectively.

Method		AU								Avg.
		1	2	4	6	9	12	25	26	
Static image-based	EAC-Net [19]	41.5	26.4	66.4	50.7	**80.5**	**89.3**	88.9	15.6	48.5
	JAA-Net [35]	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
	ARL [36]	43.9	42.1	63.6	41.8	40.0	76.2	95.2	66.8	58.7
	SMA-Net [20]	53.4	54.2	64.0	57.0	47.0	76.6	92.0	55.2	64.2
Static AU relationship modeling	SRERL [17]	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
	FAUDT [13]	46.1	48.6	72.8	56.7	50.0	72.1	90.8	55.4	61.5
	FAN-Trans [57]	56.4	50.2	68.6	49.2	57.6	75.6	93.6	58.8	63.8
	ME-GraphAU[29]	54.6	47.1	72.9	54.0	55.7	76.7	91.1	53.0	63.1
Spatio-temporal	STRAL[37]	52.2	47.4	68.9	47.8	56.7	72.5	91.3	67.6	63.0
	HSTR-Net[47]	54.3	50.8	70.1	**66.6**	59.6	68.0	**97.9**	**69.8**	62.9
	KS [21]	53.8	59.9	69.2	54.2	50.8	75.8	92.2	46.8	62.8
	WSRTL [52]	57.3	51.8	74.3	49.8	44.8	79.3	94.6	64.6	64.6
	Ours (ResNet-50)	61.4	57.7	70.9	57.1	48.3	75.7	91.5	56.7	64.9
	Ours (Swin-B)	**65.4**	**60.2**	**75.2**	50.2	52.4	74.3	93.7	58.2	**66.2**

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

where p_n^t and y_n^t are the n_{th} AU's prediction and the corresponding ground-truth for the frame f^t, respectively; a w_n is calculated for each AU based on the training set to alleviate label imbalance issue; the p_n^t at the beginning of the second term $p_n^t(1-y_n^t)\log(1-p_n^t)$ dynamically down-weights the contribution of negative samples (inactive AUs), as inactive AUs significantly outnumber active ones in the training set. Besides, a cross-entropy loss is utilized to individually supervise regional AU combination predictions of all frames produced by the first stage of the HSR module as:

$$\mathcal{L}_{\text{sub}} = \sum_{t=1}^{T} \sum_{\text{sub}=\{\text{up,mid,down}\}} \text{CE}(P_{\text{sub}}^t, Y_{\text{sub}}^t) \qquad (11)$$

where P_{sub}^t and Y_{sub}^t denote the AU combination prediction and the corresponding ground-truth of a facial region in the frame f^t and CE denotes the cross-entropy function. By predicting such AU combinations consisting of multiple AUs located in the same facial region, the network is encouraged to model underlying dependencies among spatially adjacent AUs in each facial region. Consequently, the overall loss function for training the proposed network combines the two loss functions described above as:

$$\mathcal{L} = \mathcal{L}_{\text{AU}} + \lambda \mathcal{L}_{\text{sub}} \qquad (12)$$

where λ balances the contribution of the two losses.

4. Experiments

4.1. Experimental setup

Datasets: Our MDHR is evaluated on two AU recognition benchmark datasets: BP4D [59] and DISFA [31]. BP4D is

made up of 328 facial videos containing around 140,000 frames collected from 23 females and 18 males. Meanwhile, DISFA contains 27 facial image sequence (totally $130,815$ frames) recorded from 12 females and 15 males who were asked to watch Youtube videos. Each frame in BP4D and DISFA is annotated with occurrence labels corresponding to 12 and 8 AUs, respectively.

Implementation details: We follow previous approaches [38, 60] to apply MTCNN [58] to crop and align a 224×224 face region from each frame, and conduct subject-independent three-folds cross-validation for each dataset, where the reported results are achieved by averaging the validation results of three folds. We pad k frames that same to the first frame / last frame at the beginning / end of each face video to ensure all frames can be processed by our model. AdamW [28] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ is employed for training and the λ in Eq. 12 is set to 0.01. A cosine decay learning rate scheduler is utilized, with an initial value of 0.0001. Both backbones are pre-trained on ImageNet [9]. All our experiments are conducted using NVIDIA A100 GPUs based on the open-source PyTorch library. More detailed model, training/validation, and dataset settings are provided in the Supplementary Material.

Metrics: Following previous AU recognition studies [5, 22, 38], a common metric: frame-based F1 score ($F1 = 2\frac{P \cdot R}{P+R}$), is employed to evaluate the performance of our MDHR, which takes both recognition precision P and recall rate R into consideration.

4.2. Comparison with state-of-the-arts

Table 1 and Table 2 compare our approach with previous state-of-the-art AU recognition methods, including eight static image-based methods [13, 17, 19, 20, 29, 35, 36, 57] (where four methods specifically conduct AU relationship modeling) and four spatio-temporal methods [21, 37, 47, 52]. It can be observed that our MDHR achieved new SOTA results on both datasets, with F1-scores of 66.6% (ResNet50 backbone) and 66.2% (Swin-Transformer backbone) on BP4D and DISFA, respectively. Particularly, it has clear advantages over all static image-based methods, e.g., outperformed previous state-of-the-art static AU relationship modelling method [29] with 1.1% (ResNet) and 0.6% (Swin-Transformer) improvements on BP4D, as well as 1.8% (ResNet) and 3.1% (Swin-Transformer) improvements on DISFA, respectively. Meanwhile, our approach is also superior to previous spatio-temporal methods, achieving 0.7% and 1.6% higher F1 results over the best model WSRTL [52] on BP4D and DISFA, respectively.

These results suggest that: (i) the proposed MDHR is effective and robust in AU recognition, as it achieved both best and the second best performances on both datasets under two backbone settings; (ii) jointly modelling spatio-temporal relationship among AUs could lead to additional performance gains compared to approaches [29, 46] that only consider their spatial relationship; and (iii) our MDHR can better capture AU-related spatio-temporal cues over existing spatio-temporal AU recognition approaches [47, 52]. We didn't compare approaches that utilized additional face datasets to train AU models [56] despite our MDHR still clearly outperformed them.

Backbone	MFD	HSR			TCN	F1-score
		AFE	Aux	CRM		
✓						63.3
✓	✓					64.6
✓		✓				64.1
✓		✓	✓			64.5
✓		✓	✓	✓		65.1
✓	✓	✓				65.3
✓	✓	✓	✓			65.7
✓	✓	✓	✓	✓		66.3
✓	✓	✓	✓	✓	✓	**66.6**

Table 3. Average AU recognition F1 scores (%) achieved by various settings using ResNet50 backbone on BP4D dataset, where **AFE**, **Aux**, and **CRM** denote the AU-specific feature extractor, the added auxiliary branch and the Cross-regional AU relationship modelling block, all of which belong to the HSR module.

4.3. Ablation studies

We perform ablation studies on BP4D dataset to demonstrate various aspects of our approach, where the default setting employs the ResNet as the backbone and asymmetric loss (Eqa. 10) for the model training. We further provide more ablation results (e.g., the influence of the number of adjacent frames k, statistical analysis, model complexity analysis, etc.) in the Supplementary Material.

Contribution of each component: Table 3 compares contributions of different modules. Firstly, our MFD module brought 1.3% absolute improvement, highlighting the effectiveness of the MFD in capture AU-related spatio-temporal facial behaviour cues. Meanwhile, individually employing the HSR module boosting the F1 score from 63.3% to 65.1%, validating the importance of modelling hierarchical spatio-temporal relationship among AUs. Specifically, the use of AU-specific feature extractors to individually learn each AU from its sliced facial region improved the F1 score from 63.3% to 64.1%, and the auxiliary branch also contributes additional 0.4% improvement. Finally, we found that combining our MFD and HSR module with the TCN resulted in the best performance, which validates that these two modules can learn complementary AU-related cues to further enhance AU recognition performance.

Analysis of the MFD module: Table 4 investigates our MFD module based on the system (baseline) that combines the backbone and AU-specific feature extractors. It is clear that even using facial dynamic learned from the outputs of a single and two backbone layers can consistently benefit

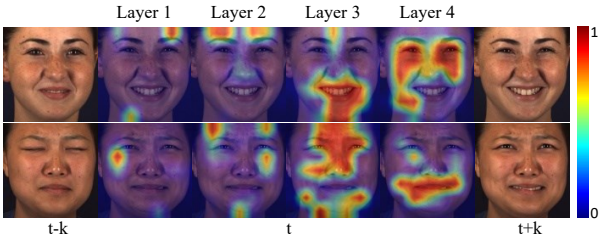

Figure 3. Visualization of adaptive weight matrices learned by the MFD module. The weight matrices learned for feature maps of shallow layers (layer 1 and 2) emphasized subtle motions (e.g., subtle eyebrow and cheek motions), while large check and mouth movements are captured in deeper layers (layer 3 and 4).

Method	baseline	Alone		Combination			Sum	Cat	AW
layer 1		✓		✓		✓	✓	✓	✓
layer 2				✓		✓	✓	✓	✓
layer 3					✓	✓	✓	✓	✓
layer 4			✓		✓		✓	✓	✓
F1-score	64.1	64.3	64.7	64.6	64.8	64.7	64.9	64.9	**65.3**

Table 4. Results of different MFD module settings, where the left part displays results achieved by computing facial dynamic at different layers and their combinations (combined using our adaptive weighting), while the right side displays results of two other fusion strategy applied to combine facial dynamic of all scales, and AW denotes adaptive weighting.

the recognition, while combining dynamic of all scales resulted in the largest improvement. This suggests that facial dynamic extracted by our MFD at different spatial scales contain complementary and useful cues for AU recognition., i.e., our MFD can emphasize each AU-related cues at its most related spatial scales (illustrated in Figure 3). Although simply adding or concatenating unweighted dynamic features of all spatial scales can already lead to performance gains, our adaptive weighting block still show clear advantage over them, suggesting that it can effectively consider the importance of each spatial scale on different AUs' recognition.

AU relationship modeling method	F1-score
Fully-connected	64.6
Aux + Locally connected	64.2
Aux + Cross-regional fully-connected	64.9
Aux + Fully connected	64.8
Aux + Connecting cross-regional activated AUs	**65.1**

Table 5. Results of different edge connection strategies.

Analysis of the HSR module: Table 3 first demonstrates that not only the HSR module brought clear improvements but also both of its local and cross-regional AU relationship modelling blocks can improve AU predictions, i.e., all its block (e.g., AFE, Aux and CRM) positively contributed to the final performance. Figure 4 further visualizes the

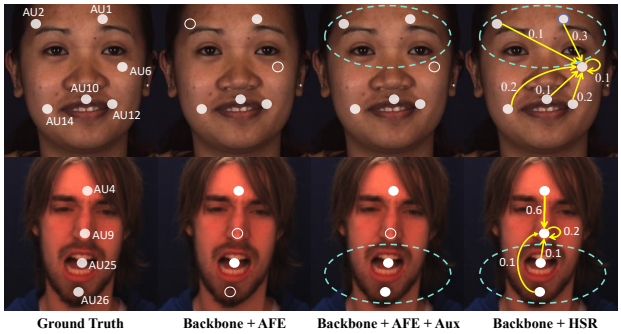

Figure 4. Visualization of AU predictions under three HSR settings, where white solid and hollow dots denote activated and inactivated AUs. The green dotted circles denote the local AU relationship modelling, while the yellow lines/weights denote the graph edges describing the association between AUs. It can be observed that the local relationship modelling can effectively model dependencies between AUs in the same region to make better predictions (e.g., AU2 and AU26 in column 3), while additionally use cross-regional AU relationship modelling can further utilize the learned relationship cues to improve AU predictions in different facial regions (e.g., AU6 and AU9 in column 4).

impact of these blocks. Additionally, Table 5 compares different AU graph edge connection settings of the cross-regional AU relationship modelling block, where the setting that connects each activated AU to all other AUs located in different facial region achieved the superior performance to other edge settings, i.e., this setting can effectively model cross-regional AU relationship. Importantly, the HSR is not sensitive to different edge connection settings when cross-regional AU relationship is considered.

5. Conclusion

This paper proposes a novel MDHR that not only computes facial dynamics at different spatial scales as AUs could exhibit heterogeneity in their ranges and magnitudes, but also models hierarchical spatio-temporal relationships among AUs. Results show that the proposed two modules can effective capture AU-related dynamics and their relationships, making our MDHR becoming the new SOTA AU recognition method. The main limitations are that our facial region slicing strategy could be potentially improved and more advanced graph edge learning strategies could be applied to HSR for better modelling relationships.

6. Acknowledgement

The work is supported by National Natural Science Foundation of China under Grant 82261138629; Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515010688 and Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20220531101412030.

References

[1] Luigi Cattaneo and Giovanni Pavesi. The facial motor system. *Neuroscience & Biobehavioral Reviews*, 38:135–159, 2014. 2, 3, 5

[2] Yanan Chang and Shangfei Wang. Knowledge-driven self-supervised representation learning for facial action unit recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20417–20426, 2022. 2

[3] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F. Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 25–32, 2017. 2

[4] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 25–32. IEEE, 2017. 2

[5] Nikhil Churamani, Sinan Kalkan, and Hatice Gunes. Aulacaps: Lifecycle-aware capsule networks for spatio-temporal analysis of facial actions. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021. 2, 7

[6] 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. 2

[7] Zijun Cui, Tengfei Song, Yuru Wang, and Qiang Ji. Knowledge augmented deep neural networks for joint facial expression and action unit recognition. *Advances in Neural Information Processing Systems*, 33:14338–14349, 2020. 1

[8] Zijun Cui, Chenyi Kuang, Tian Gao, Kartik Talamadupula, and Qiang Ji. Biomechanics-guided facial action unit detection through force modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8694–8703, 2023. 1

[9] 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. 7

[10] Joy O Egede, Siyang Song, Temitayo A Olugbade, Chongyang Wang, C De C Amanda, Hongying Meng, Min Aung, Nicholas D Lane, Michel Valstar, and Nadia Bianchi-Berthouze. Emopain challenge 2020: Multimodal pain evaluation from facial and bodily expressions. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 849–856. IEEE, 2020. 1

[11] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 1

[12] Shuangjiang He, Huijuan Zhao, Jing Juan, Zhe Dong, and Zhi Tao. Optical flow fusion synthesis based on adversarial learning from videos for facial action unit detection. In *The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021)*, pages 561–571. Springer, 2022. 2, 3

[13] 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. 1, 2, 6, 7

[14] Shashank Jaiswal and Michel Valstar. Deep learning the dynamic appearance and shape of facial action units. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8, 2016. 1, 2

[15] Shashank Jaiswal and Michel Valstar. Deep learning the dynamic appearance and shape of facial action units. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–8. IEEE, 2016. 2

[16] Zhao Kaili, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016. 1

[17] 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*, pages 8594–8601, 2019. 2, 5, 6, 7

[18] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1841–1850, 2017. 1, 2

[19] 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. 2, 5, 6, 7

[20] Xiaotian Li, Zhihua Li, Huiyuan Yang, Geran Zhao, and Lijun Yin. Your "attention" deserves attention: A self-diversified multi-channel attention for facial action analysis. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08, 2021. 2, 6, 7

[21] Xiaotian Li, Xiang Zhang, Taoyue Wang, and Lijun Yin. Knowledge-spreader: Learning semi-supervised facial action dynamics by consistifying knowledge granularity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20979–20989, 2023. 3, 6, 7

[22] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer vision and pattern recognition*, pages 10924–10933, 2019. 3, 7

[23] Zhihua Li, Xiang Deng, Xiaotian Li, and Lijun Yin. Integrating semantic and temporal relationships in facial action unit detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5519–5527, 2021. 2

[24] Zhihua Li, Zheng Zhang, and Lijun Yin. Sat-net: Self-attention and temporal fusion for facial action unit detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5036–5043, 2021. 2

[25] Songtao Liu, Di Huang, and Yunhong Wang. Learning spatial fusion for single-shot object detection. *arXiv preprint arXiv:1911.09516*, 2019. 4

[26] Zhilei Liu, Jiahui Dong, Cuicui Zhang, Longbiao Wang, and Jianwu Dang. Relation modeling with graph convolutional networks for facial action unit detection. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 489–501. Springer, 2020. 2

[27] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. Teinet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11669–11676, 2020. 3

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 7

[29] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*, 2022. 1, 2, 3, 5, 6, 7

[30] 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. 1

[31] 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. 2, 6

[32] 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. 1, 2

[33] Tao Pu, Tianshui Chen, Yuan Xie, Hefeng Wu, and Liang Lin. Au-expression knowledge constrained representation learning for facial expression recognition. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 11154–11161. IEEE, 2021. 1

[34] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pages 3–12, 2019. 1

[35] 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. 6, 7

[36] 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*, 13(3):1274–1289, 2019. 1, 2, 6, 7

[37] Zhiwen Shao, Lixin Zou, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Spatio-temporal relation and attention learning for facial action unit detection. *arXiv preprint arXiv:2001.01168*, 2020. 2, 6, 7

[38] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaanet: joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129:321–340, 2021. 1, 2, 7

[39] Zhiwen Shao, Yong Zhou, Hancheng Zhu, Wen-Liang Du, Rui Yao, and Hao Chen. Facial action unit recognition by prior and adaptive attention. *Electronics*, 11(19):3047, 2022. 2

[40] Zhiwen Shao, Yong Zhou, Jianfei Cai, Hancheng Zhu, and Rui Yao. Facial action unit detection via adaptive attention and relation. *IEEE Transactions on Image Processing*, 32: 3354–3366, 2023. 2

[41] Zhiwen Shao, Yong Zhou, Jianfei Cai, Hancheng Zhu, and Rui Yao. Facial action unit detection via adaptive attention and relation. *IEEE Transactions on Image Processing*, 2023. 1

[42] Juan Song and Zhilei Liu. Self-supervised facial action unit detection with region and relation learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[43] Siyang Song, Enrique Sánchez-Lozano, Mani Kumar Tellamekala, Linlin Shen, Alan Johnston, and Michel Valstar. Dynamic facial models for video-based dimensional affect estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 3

[44] Siyang Song, Shashank Jaiswal, Linlin Shen, and Michel Valstar. Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing*, 13(2):829–844, 2020. 1

[45] Siyang Song, Yuxin Song, Cheng Luo, Zhiyuan Song, Selim Kuzucu, Xi Jia, Zhijiang Guo, Weicheng Xie, Linlin Shen, and Hatice Gunes. Gratis: Deep learning graph representation with task-specific topology and multi-dimensional edge features. *arXiv preprint arXiv:2211.12482*, 2022. 1

[46] 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*, pages 5993–6001, 2021. 1, 2, 7

[47] Wenyu Song, Shuze Shi, Yu Dong, and Gaoyun An. Heterogeneous spatio-temporal relation learning network for facial action unit detection. *Pattern Recognition Letters*, 164:268–275, 2022. 2, 3, 6, 7

[48] Yan Tong, Wenhui Liao, and Qiang Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1683–1699, 2007. 1

[49] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 5

[50] Zihan Wang, Siyang Song, Cheng Luo, Yuzhi Zhou, Weicheng Xie, Linlin Shen, et al. Spatio-temporal au relational graph representation learning for facial action units detection. *arXiv preprint arXiv:2303.10644*, 2023. 2

[51] Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. Au-assisted graph attention convolutional network for micro-expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2871–2880, 2020. 1

[52] Jingwei Yan, Jingjing Wang, Qiang Li, Chunmao Wang, and Shiliang Pu. Weakly supervised regional and temporal learning for facial action unit recognition. *IEEE Transactions on Multimedia*, 2022. 3, 6, 7

[53] Jingwei Yan, Jingjing Wang, Qiang Li, Chunmao Wang, and Shiliang Pu. Weakly supervised regional and temporal learning for facial action unit recognition. *IEEE Transactions on Multimedia*, 25:1760–1772, 2023. 3

[54] Huiyuan Yang and Lijun Yin. Learning temporal information from a single image for au detection. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. 3

[55] 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. 1

[56] Jing Yang, Yordan Hristov, Jie Shen, Yiming Lin, and Maja Pantic. Toward robust facial action units' detection. *Proceedings of the IEEE*, 2023. 2, 3, 7

[57] Jing Yang, Jie Shen, Yiming Lin, Yordan Hristov, and Maja Pantic. Fan-trans: Online knowledge distillation for facial action unit detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6019–6027, 2023. 1, 2, 6, 7

[58] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23 (10):1499–1503, 2016. 7

[59] 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. 2, 6

[60] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Classifier learning with prior probabilities for facial action unit recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 7

[61] 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. 2