Hybrid Message Passing with Performance-Driven Structures for Facial Action Unit Detection

Tengfei Song[1], Zijun Cui[2], Wenming Zheng[1], and Qiang Ji[2]
[1]Southeast University
[2]Rensselaer Polytechnic Institute
songtf@seu.edu.cn, cuiz3@rpi.edu, wenming_zheng@seu.edu.cn, qji@ecse.rpi.edu

Abstract

Message passing neural network has been an effective method to represent dependencies among nodes by propagating messages. However, most of message passing algorithms focus on one structure and messages are estimated by one single approach. For real-world data, like facial action units (AUs), the dependencies may vary in terms of different expressions and individuals. In this paper, we propose a novel hybrid message passing neural network with performance-driven structures (HMP-PS), which combines complementary message passing methods and captures more possible structures in a Bayesian manner. Particularly, a performance-driven Monte Carlo Markov Chain sampling method is proposed for generating high performance graph structures. Besides, hybrid message passing is proposed to combine different types of messages, which provide the complementary information. The contribution of each type of message is adaptively adjusted along with different inputs. The experiments on two widely used benchmark datasets, i.e., BP4D and DISFA, validate that our proposed method can achieve the state-of-the-art performance.

1. Introduction

Facial action units (AUs) detection is an essential technique to analyze human expressions in the field of artificial intelligence. Ekman proposed Facial Action Coding System (FACS) [10] to model the relation between facial muscle movements, $i.e.$, AUs, and facial expressions. Therefore, more studies estimate the human emotions and facial behaviors via facial AU detection.

Convolutional neural networks (CNN) [15, 27, 41, 26, 44] have been extensively applied to various classification and regression tasks because of the powerful local feature representation capabilities. However, for facial AU detection, the structural information and dependencies among different AUs [17] should be taken into consideration during AU detection. CNN can not fully characterize the relationships among different AUs. Recently, graph neural networks [13, 7, 9], also called message passing neural networks (MPNN), have been popular algorithm to characterize the objects and their dependencies. Recent studies [23, 11] employ message passing neural networks to solve facial AU related tasks.

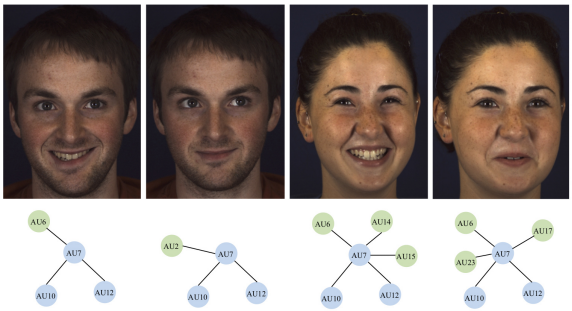


Figure 1. Some facial images and AU occurrences for happiness expression. The AUs with blue nodes are common AU occurrences and the AUs with green nodes are different AU occurrences. Even for the same expression, the AU occurrences and dependencies may be different. For example, AU6 (Cheek Raiser) occurs in the first image and does not occur in the second image.

Generally, previous studies [17, 38] utilized prior knowledge from FACS or training labels to define the graph structure so as to represent the dependencies among AUs. These common dependencies among AUs can provide useful structural information. However, one type of constant relationship among AUs is not sufficient to fully characterize the sophisticated correlations among AUs. The dependencies among AUs may vary given different expressions. Besides, individual difference is also a factor to induce different correlations among AUs. Even for the same expression, as shown in Figure 1, AU occurrences can be different. Therefore, different graph structures should be taken into consideration for effectively capturing different types of dependencies for AU detection.

For the graph structure of MPNN, some studies focused on gradient-based methods to learn the discrete structure by bilevel training [12]. However, the gradient-based is easy to fall into the local optimal solution. Another popular graph structure learning strategy is sampling-based method. Metropolis-Hasting Monte Carlo Markov Chain sampling (MH-MCMC) [14] is commonly applied to structure learning in probabilistic graphical model, which can sample an ensemble of graph structures with the goodness of fit to data. To improve the effectiveness of message passing neural networks, more effective ways to generate graph structures from data are necessary to be exploited.

Although there are many different ways to define messages, most of message passing neural networks [13, 39] propagate information by only providing one type of message. For facial AUs, some AUs are positively correlated, like AU4 (brow lower) and AU7 (lid tightener), and some AUs are negatively correlated, like AU12 (lip corner puller) and AU15 (lip corner depressor). Apparently, traditional message passing algorithms can not fully characterize these different relationships among AUs. More message categories and more effective ways to combine different messages should be provided to represent diverse dependencies and propagate the complementary information.

Based on aforementioned considerations, in this paper, we propose a novel hybrid message passing neural network with performance-driven structures (HMP-PS) for facial AU detection, which generates the predictions by considering different graph structures in a Bayesian manner. For the structure of MPNN, traditional MH-MCMC sampling draws the distribution of graph structure based on data likelihood. Different from MH-MCMC, we focus on sampling graph structures based on their performances on facial AU detection. Therefore, we propose performance-driven Monte Carlo Markov Chain sampling (P-MCMC) to generate the graph structures with high performance. And then, the graph structures with high performance can be used for MPNN. Besides, we propose the hybrid message passing to dynamically combine three types of messages such that more complementary information can be dynamically propagated among different nodes.

The contributions of this paper can be summarized as follows:

- We propose performance-driven Monte Carlo Markov Chain sampling to generate high performance graph structures.
- We propose the hybrid message passing to dynamically combine different types of messages to exploit their complementary information.
- Our method achieves better performance than the state-of-the-art methods on two widely used benchmark datasets, $i.e.$, BP4D and DISFA.

2. Related Works

2.1. Facial Action Unit Detection

With high performance computing and large scale datasets, CNN has shown great power in automatic facial action unit detection by extracting effective appearance features [18, 27, 28, 19, 5]. As defined in FACS [10], there are important dependencies among different AUs such that more researches focus on incorporating the AU dependencies to CNN model for facial AU detection.

The semantic dependencies can be characterized by various approaches. Walecki $et\ al.$ [36] employed a conditional random field (CRF) model to exploit the relations among different AUs. A combination between CRF and deep learning is applied and the parameters are learned jointly. Zhao $et\ al.$ [45] proposed a deep model to simultaneously capture the salient regions and exploit the AU dependencies. The improvement validates the effectiveness of encoding the AU dependencies. Cui $et\ al.$ [6] employed a Bayesian Network to encode the relations between AU labels, which were applied to correct the noisy AU labels. Recently, more studies employed graph neural network to capture the AU dependencies and provide a discriminative graph representation. Li $et\ al.$ [17] learned the graph based on FACS and incorporated the learned graph as prior information to graph neural network for AU detection. Combined with prior knowledge, graph neural network provides more effective way to capture the dependencies among AUs. Niu $et\ al.$ [23] calculated the conditional probabilities of AU occurrences from AU labels as the weighted adjacency matrix for graph convolution, which validated that the prior information about AU dependency is helpful for semi-supervised AU detection. Most of these studies learn a constant graph to model the relations among AUs and the individual and dynamic dependencies among AUs are not addressed. In this paper, we provide a more flexible and complementary way to represent the dynamic relations among AUs.

2.2. Message Passing Neural Network

Recently, message passing neural networks (MPNNs), also called graph neural networks (GNNs), have been popular methods to characterize different objects and their relationships [39, 32, 33, 30, 29, 31]. Many researches work on graph structure learning and the approach to calculate the messages for improving the efficiency of MPNN.

For MPNNs, effective graph structure is essential to propagate messages in right directions. Generally, there are two strategies for structure learning. One is sampling-based method and another is gradient-based method. For gradient-based method, the structure is discrete and not differentiable. Some studies [40, 22, 2] employed the Gumbel-Sigmoid functions [16] to relax the discrete search to be continuous such that the gradient back-propagate can be

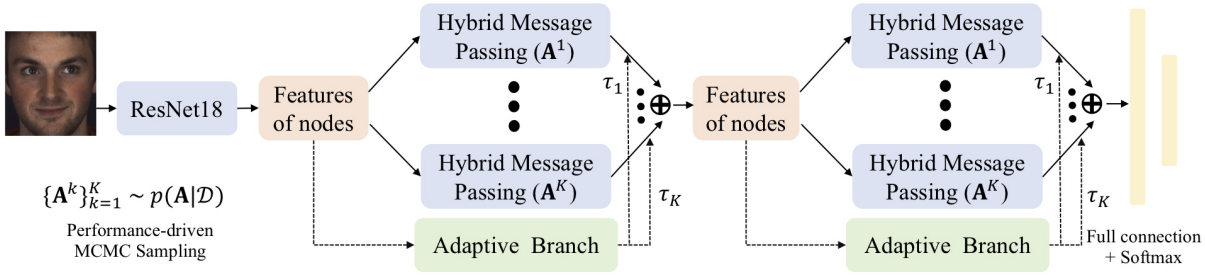


Figure 2. The architecture of hybrid message passing neural network with performance-driven structures for Facial AU detection. ResNet18 is the backbone network. $\mathbf{A}^1, ..., \mathbf{A}^K$ denote different graph structures generated by performance-driven MCMC sampling.

conducted. Although gradient-based method needs small computational power, it tends to fall into local minimum solution. More importantly, it produces only one structure, hence tends to overfit. Sampling-based methods are generally used for Bayesian structure learning. Especially, Metropolis-Hasting MCMC sampling is popular method to estimate the posterior distribution of graph structure [14, 34]. It produces an ensemble of graph structures and can generalize better. However, most of these sampling methods try to sample the structures based on data likelihood. For facial AU detection, a sampling method to improve the performance of MPNN should be taken into consideration.

The main idea of MPNN is to iteratively update the hidden state at each node by aggregating received messages. Gilmer *et al.* [13] provided a review of some GNN variants and unified them into MPNN. Previous studies have developed various methods to calculate messages and update the hidden state at each node. Duvenaud *et al.* [9] proposed to separately sum over connected nodes and connected edges. However, it is problematic to identify correlations between node states and edge states. Li *et al.* [20] proposed to aggregate neighbor nodes based on discrete adjacency matrix and update hidden state for each node by Gated Recurrent Unit [3]. Battagelia *et al.* [1] employed a neural network to calculate messages with the concatenation of node hidden states and edge hidden states as input. The update function is also a neural work that takes hidden states and messages as input. Defferrard *et al.* [7] employed the parameterized weights by the eigenvectors of graph Laplacian L to calculate the messages. The sigmoid activation function and the ReLU function were provided as update functions. Most of these previous works employed one type of message and the constant connections, which can not reflect the sophisticated correlations among AUs in real-world data. Li *et al.* [42] proposed a dynamic message passing network that adaptively generates the weighted adjacency matrix for message passing. However, one type of message is still limited to representing all useful information from neighbor nodes. To provide more effective way to aggregate messages, we exploit more types of messages and dynamically combine these messages.

3. Method

The overall architecture of the proposed method is presented in Figure 2. We use adjacency matrix $\mathbf{A}$ to represent the graph structure. Inspired by Bayesian inference, deep features are learned by considering different possible dependencies among AUs in a Bayesian manner. Particularly, different structures are sampled via the proposed performance-driven MCMC sampling approach, based on which we perform the hybrid message passing. The final output is obtained through the integration over all the sampled candidate structures. Our method is formulated as Bayesian inference, *i.e.*,

$$\begin{aligned} p(h|\mathcal{D}) &= \int p(h|\mathbf{A},\mathcal{D})p(\mathbf{A}|\mathcal{D})d\mathbf{A} \\ &\approx \sum_{k=1}^{K} p(h|\mathbf{A}^k,\mathcal{D}), \quad \mathbf{A}^k \sim p(\mathbf{A}|\mathcal{D}) \\ &= \sum_{k=1}^{K} f_{\text{HMP}}(\mathcal{D},\mathbf{A}^k), \end{aligned} \tag{1}$$

where h is the output of our message passing method, *i.e.*, node feature, $\mathcal{D}$ is the data, $\mathbf{A}^k$ is the k-th sample generated from $p(\mathbf{A}|\mathcal{D})$ and f_{HMP} is the hybrid message passing. More possible structures from the posterior distribution of graph should be considered for message passing. Below, we first introduce performance-driven MCMC sampling to generate structure samples, and then the hybrid message passing algorithm, *i.e.*, f_{HMP}, finally hybrid message passing with performance-driven sampled structures.

3.1. Performance-Driven MCMC Sampling for Discrete Structure Learning

As shown in Eq.(1), we first need to obtain the posterior distribution of graph such that multiple graph structures can be considered. Directly estimating the posterior $p(\mathbf{A}|\mathcal{D})$ is intractable and there is no analytic solution. For structure learning, Metropolis Hasting Monte Carlo Markov Chain sampling (MH-MCMC) [14] can be applied to sample an ensemble of graph structures to approximate posterior distribution of graph structure, *i.e.*, $p(\mathbf{A}|\mathcal{D})$. $p(\mathbf{A}|\mathcal{D})$ can be represented as $p(\mathbf{A}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{A})p(\mathbf{A})}{p(\mathcal{D})}$. Given an uniform prior $p(\mathbf{A})$, $p(\mathbf{A}|\mathcal{D})$ is proportional to $p(\mathcal{D}|\mathbf{A})$, and thus

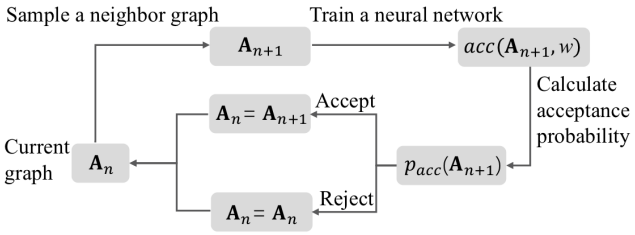


Figure 3. The framework of performance-driven MCMC sampling method to generate the graph structures.

thus structures can be samples based on data likelihood $p(\mathcal{D}|\mathbf{A})$ to approximate $p(\mathbf{A}|\mathcal{D})$. The sampled structures have high data likelihood and characterize the goodness of fit to data.

We use adjacency matrix, *i.e.*, $\mathbf{A} \in \mathbb{R}^{N\times N}$ to represent graph structure. N is the number of nodes. If there is an edge from node i to node j, $\mathbf{A}_{i,j}$ should be 1, otherwise 0. $\mathbf{A}$ is symmetric for undirected graph and is asymmetric for directed graph or directed acyclic graph. In MH-MCMC, given current sample $\mathbf{A}_n$, a proposal probability of $\mathbf{A}_{n+1}$ is required for generating the next sample $\mathbf{A}_{n+1}$. and the proposal probability is defined as

$$p_{pro}(\mathbf{A}_{n+1}) = \frac{1}{\mathcal{N}(\mathbf{A}_n)}, \tag{2}$$

in which $\mathcal{N}(\mathbf{A}_n)$ denotes the number of neighbor graph of $\mathbf{A}_n$. Particularly, the neighbor graph of $\mathbf{A}_n$ is obtained by revising one edge of $\mathbf{A}_n$, which means the neighbor graph of $\mathbf{A}_n$ only contains one different edge compared to $\mathbf{A}_n$. Given the proposed graph sampled from the proposal probability, the acceptance probability is defined to decide whether to accept this new graph. The acceptance probability can be calculated by

$$p_{acc}(\mathbf{A}_{n+1}) = min\{1, \frac{p(\mathcal{D}|\mathbf{A}_{n+1})}{p(\mathcal{D}|\mathbf{A}_n)} \cdot \frac{p_{pro}(\mathbf{A}_{n+1})}{p_{pro}(\mathbf{A}_n)}\}, \tag{3}$$

in which $p(\mathcal{D}|\mathbf{A}_{n+1})$ is the likelihood of $\mathbf{A}_{n+1}$. Assuming uniform $p(\mathbf{A})$, the posterior $p(\mathbf{A}|\mathcal{D})$ is proportional to the likelihood $p(\mathcal{D}|\mathbf{A})$ such that the sampled graphs can draw the distribution of posterior $p(\mathbf{A}|\mathcal{D})$ based on the likelihood $p(\mathcal{D}|\mathbf{A})$. A new graph with higher data likelihood has higher probability to be accepted.

For classification tasks, we want to sample the graph structures with higher classification accuracy. Previous MH-MCMC sampling is based on the likelihood of graph to generate samples. We propose performance-driven MCMC sampling method to conduct the sampling process based on the performance of neural network such that we can ensure the next sampled structure has high probability to outperform the current one. The accept probability is defined as

$$p_{acc}(\mathbf{A}_{n+1}) = min\{1, \frac{acc(\mathbf{A}_{n+1}, w)}{acc(\mathbf{A}_n, w)} \cdot \frac{p_{pro}(\mathbf{A}_{n+1})}{p_{pro}(\mathbf{A}_n)}\}, \tag{4}$$

in which $acc(\mathbf{A}_{n+1}, w)$ is the classification accuracy of neural network and w denotes the parameter of neural network. As shown in Figure 3, we can iteratively generate a group of graph structures, which draw the distribution of the performance of neural network. Multiple graph structures with high classification performance can be selected and provide different types of dependencies among nodes. For AU detection, we employ the F1 score to estimate the performance. The graph structures with top K highest F1 score, *i.e.*, $\{\mathbf{A}^k\}_{k=1}^K$ are used as the candidate graph structures for MPNN such that more possible dependencies among nodes can be captured.

3.2. Hybrid Message Passing

In Eq.(1), we try to develop a more effective message passing function to improve the performance. MPNN iteratively updates the node features by propagating information between neighbor nodes, which provides the graph representation. Generally, given the node features, we can use message functions and update functions to characterize the message passing neural network as

$$\begin{cases} m_i^l = \mathcal{F}(h_i^l, \{h_j^l | j \in \mathcal{N}(i)\}) \\ h_i^{l+1} = \mathcal{U}(h_i^l, m_i^l), \end{cases} \tag{5}$$

in which h_i^l is the node feature for node i in the l-th layer, m_i^l denotes the total message that node i received, $\mathcal{N}(i)$ denotes the neighbor nodes of node i, $\mathcal{F}$ is the message function and $\mathcal{U}$ is the update function.

Although previous studies about MPNNs [13, 9, 1] proposed different methods to calculate messages, most of them employ one type of message to propagate the information. For real-world data, it is not sufficient to consider only one type of message. Especially for facial AUs, some are positive related and some are negative related. Besides, the individual differences may also induce the diversity of AU dependencies. Therefore, it is necessary to combine different types of messages and provide more complementary information for message passing. As shown in Figure 4, we propose the hybrid message passing (HMP) to dynamically combine different messages. If there is a pathway from node j to node i, given the node features (hidden states), *i.e.*, h_i and $h_j \in \mathbb{R}^{1\times d}$, for node i and node j, we define three types of messages as

$$\begin{cases} m_{i,j}^{ls} = \text{MLP}_s(h_i^l + h_j^l) \\ m_{i,j}^{lc} = \text{MLP}_c(\text{cat}[h_i^l, h_j^l]) \\ m_{i,j}^{ld} = \text{MLP}_d(h_i^l - h_j^l), \end{cases} \tag{6}$$

in which m^{ls} denotes the summation message for the l-th layer, m^{lc} denotes the concatenation message and m^{ld} denotes the differential message. MLP_s, MLP_c and MLP_d are different Multilayer Perceptrons (MLP) to estimate different messages. Different types of messages can reflect different dependencies, which provide a potential way to exploit more complementary information. Rather than providing equal or constant weights for different messages, we

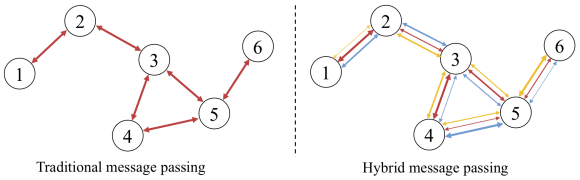


Figure 4. **Left:** Traditional message passing with one type of message. **Right:** Hybrid massage passing with three types of messages. Different colors of arrows denote different messages and the thickness of arrow denotes the value of dynamic weight.

propose to generate the dynamic weights to adaptively select important messages and suppress the useless messages. Inspired by attention mechanism [35, 37], for each pathway, we employ an additional branch to generate the dynamic weights to combine different messages. The dynamic weights can be calculated by

$$p_{ij}^l = \text{softmax}(\text{Cat}[h_i^l, h_j^l]\text{W} + \text{b}), \tag{7}$$

in which $p_{ij}^l \in \mathbb{R}^{1\times 3}$ is a vector to represent the weights for three types of messages. $\text{W} \in \mathbb{R}^{2d\times 3}$ and $\text{b} \in \mathbb{R}^{1\times 3}$ are parameters of this additional branch. The total message that node i received can be expressed as

$$\begin{aligned} m_i^l &= \mathcal{F}_{\text{HMP}}(h_i^l, \{h_j^l | j \in \mathcal{N}(i)\}) = \sum_{j\in\mathcal{N}(i)} m_{i,j}^l \\ &= \sum_{j\in\mathcal{N}(i)} p_{ij1}^l m_{i,j}^{ls} + p_{ij2}^l m_{i,j}^{lc} + p_{ij3}^l m_{i,j}^{ld}, \end{aligned} \tag{8}$$

in which $\mathcal{F}_{\text{HMP}}$ denotes the hybrid message function to calculate the combined message that node i received. The updated feature for node i can be represented as

$$h_i^{l+1} = \mathcal{U}(h_i^l, m_i^l) = \text{ReLU}(m_i^l), \tag{9}$$

in which ReLU is the update function and h^{l+1} is the updated feature for node i. After training, for different inputs, $i.e.$, the node features, the network will adaptively generate different weights to select useful messages, which provides an effective way to measure the dependencies among AUs.

3.3. Hybrid Message Passing with Performance-Driven Structures

In Eq.(1), different hybrid message passing layers with different structures are combined. Based on the structure samples from P-MCMC sampling and the hybrid message passing, we propose hybrid message passing with performance-driven structures. The hybrid message passing layer with one graph structure can be expressed as

$$h_i^{l+1} = f_{\text{HMP}}(h_i^l, \{h_j^l | j \in \mathcal{N}(i)\}, \mathbf{A}), \tag{10}$$

where f_{HMP} is the hybrid message passing layer (HMP) with one sampled structure $\mathbf{A}$ to update the feature for node i.

The proposed P-MCMC sampling can generate the structures with high performance. These graph structures represent different dependencies among AUs. To utilize these dependencies, we propose to combine different graph structures. Rather than providing equal weights, we provide the dynamic weights to measure the contributions of different graph structures. We use $\text{H}^l = [h_1^l; ...; h_N^l] \in \mathbb{R}^{N\times d}$ to define the features of all nodes for the l-th layer. Let N be the number of nodes and d is the dimension of node feature. As shown in Figure 2, we use an adaptive branch to generate the weights for different graph structures:

$$\tau^l = \text{softmax}((\text{H}^l\text{P} + \text{B})\text{Q}), \tag{11}$$

in which $\text{P} \in \mathbb{R}^{d\times 1}$, $\text{B} \in \mathbb{R}^{N\times 1}$ and $\text{Q} \in \mathbb{R}^{N\times K}$ are the parameters of neural network. $\tau^l \in \mathbb{R}^K$ is the weights in the l-th layer for different graph structures.

Specifically, the updated feature of our hybrid message passing with performance-driven structures (HMP-PS) for node i can be represented as

$$h_i^{l+1} = \sum_{k=1}^{K} \tau_k^l f_{\text{HMP}}(h_i^l, \{h_j^l | j \in \mathcal{N}(i)\}, \mathbf{A}^k), \tag{12}$$

in which τ_k^l is the k-th element of τ^l. The updated node feature is the linear combination of multiple hybrid message passing layers with different graph structures by a group of dynamic weights. For different input data, adaptive branch adaptively generates the weights to select graph structures.

3.4. Loss Function

As shown in Figure 2, we employ ResNet18 [15] as the backbone network to extract disentangled features for different AUs, which is similar with [23] (more details about AU feature generation are in supplemental material). After we obtain the disentangled features for different AUs, we employ two HMP-PS layers to extract more discriminative features. For each AU, one fully connected layer is provided to reduce the feature dimension and one softmax layer is employed to output probability of AU occurrences. The total loss can be written as

$$\mathcal{L} = \sum_{c=1}^{N} [y_c log(p(y_c)) + (1 - y_c) log(1 - p(y_c))], \tag{13}$$

in which $\mathcal{L}$ is total loss and y_c is ground truth of the c-th AU. $p(y_c)$ is predicted probability for the c-th AU occurrence.

4. Experiments

4.1. Setting

Datasets: In this paper, we evaluate our algorithm in two widely used benchmark datasets, $i.e.$, BP4D [43] and DISFA [21]. **BP4D** is a spontaneous facial AU dataset that includes 41 subjects with 18 male and 23 females. There are 8 sessions for each subject such that this dataset contains total 328 videos for 41 subjects. Specifically, about 140,000 frames are annotated with AU labels. 12 AUs are used for evaluation with the subject-exclusive three-fold cross-validation experiment protocol, which is consistent with

Table 1. The F1 score (in %) for the recognition of 12 AUs reported by the proposed HMP-PS and the state-of-the-art methods on the BP4D dataset. The best results are indicated using bold.

Methods	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg
DRML [45]	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
ROI [18]	36.2	31.6	43.4	77.1	73.7	**85.0**	87.0	62.6	45.7	58.0	38.3	37.4	56.4
DSIN [4]	51.7	40.4	**56.0**	76.1	73.5	79.9	85.4	62.7	37.3	62.9	38.8	41.6	58.9
MLCR [23]	42.4	36.9	48.1	77.5	77.6	83.6	85.8	61.0	43.7	63.2	42.1	**55.6**	59.8
JAA-Net [26]	47.2	44.0	54.9	77.5	74.6	84.0	86.5	61.9	43.6	60.3	42.7	41.9	60.0
ARL [27]	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
CMS [25]	49.1	44.1	50.3	**79.2**	74.7	80.9	**88.3**	63.9	44.4	60.3	41.4	51.2	60.6
SRERL [17]	46.9	45.3	55.6	77.1	**78.4**	83.5	87.6	60.6	**52.2**	**63.9**	47.1	53.3	62.9
LP-Net [24]	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
HMP-PS	**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**

Table 2. The F1 score (in %) for the recognition of 8 AUs reported by our HMP-PS and the state-of-the-art methods on the DISFA dataset.

Methods	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	Avg
DRML [45]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	21.8
ROI [18]	41.5	26.4	66.4	50.7	8.5	**89.3**	88.9	15.6	48.5
DSIN [4]	42.4	39.0	68.4	28.6	46.8	70.8	90.4	42.2	53.6
JAA-Net [26]	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
ARL [27]	43.9	42.1	63.6	41.8	40.0	76.2	**95.2**	66.8	58.7
CMS [25]	40.2	44.3	53.2	57.1	50.3	73.5	81.1	59.7	57.4
SRERL [17]	**45.7**	**47.8**	59.6	47.1	45.6	73.5	84.3	43.6	55.9
LP-Net [24]	29.9	24.7	**72.7**	46.8	49.6	72.9	93.8	65.0	56.9
HMP-PS	38.0	45.9	65.2	**50.9**	**50.8**	76.0	93.3	**67.6**	**61.0**

previous works [17, 26]. **DISFA** is also a spontaneous facial AU datasets with AU intensity labels. The facial images with AU intensities equal or greater than 2 are considered as the occurrence of AU. 27 videos are recorded from 12 females and 15 males and each video contains 4,845 images. In this paper, we conduct subject-exclusive three-fold cross-validation experiment protocol to evaluate our method and 8 AUs are provided for this experiment, which follows the same setting with previous studies [17, 26].

Evaluation Metrics: F1 score is provided to evaluate our algorithm. F1 score is often applied to binary classification with imbalanced data. Given the precision p and the recall r, F1 score is calculated by $\text{F1} = 2\frac{p \cdot r}{p+r}$. The F1 score of the positive class is provided for AU detection.

Implementation Details: During the node feature extraction for AUs, we employed ResNet18 [15] model pre-trained on ImageNet [8] as the initial model. And then, we use the training data to train ResNet18 to extract the AU features. For each facial image, we cropped the face region and resized the cropped image into 256×256. And then, these facial images are randomly cropped to 224×224 and the random horizontal flipping is also used for data augmentation. The dimension of each AU feature, $i.e.$, d is set to 512. The learning rate is 0.001 and the batch size is 64.

For P-MCMC sampling, the architecture of the neural network is: `Backbone` − `HMP` − `HMP` − `FC` − `Softmax`, in which we employ two hybrid message passing layers to test the performance, $i.e.$, the averaged F1 score. The graph structure is directed graph. 100 graph structures are accepted and we select the graph structures with the top-5 highest F1 scores as the candidate graph structures for HMP-PS. The learning rate to train the neural network is set to 0.001 and the batch size is set to 64.

For our HMP-PS, the architecture of neural network is: `Backbone` − `HMP-PS` − `HMP-PS` − `FC` − `Softmax` , in which employ two HMP-PS layers to predict the AU occurrence. The dimension of each type of message is set to 512 and the learning rate is set to 0.001. The batch size is set to 64. All the experiments are implemented with PyTorch on a GeForce RTX 2080 GPU.

4.2. Compared with the State-of-the-art Methods

We compare our HMP-PS algorithm with the state-of-the-art methods under the same subject-exclusive three-fold cross-validation experiment protocol on BP4D and DISFA. DRML [45], ROI [18], DSIN [4], MLCR [23], JAA-Net [26], ARL [27], CMS [25], SRERL [17] and LP-Net [24] are provided for comparison. Specifically, we focus on frame-based AU detection in this paper such that ROI-LSTM [18] and STRAL [28] are not presented as comparison. To provide a fair comparison, we directly use the results of DRML, ROI, DSIN, JAA-Net, ARL, CMS, SRERL and LP-Net reported in [45, 18, 4, 23, 26, 27, 25, 17, 24].

Table 1 provides the results of various state-of-the-art methods on BP4D. Specifically, all the comparison methods are based on deep models, which have good feature representation capabilities. However, our HMP-PS still achieves

Table 3. The F1 score (in %) of ablation experiments for the recognition of 12 AUs on the BP4D database reported by hybrid message passing neural networks with different types of graph structures. 'w/o' denotes 'without'.

Methods	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg
HMP-PS (w/o graph)	**54.1**	45.6	54.1	74.4	74.3	81.6	85.7	57.1	47.2	57.1	42.8	41.5	59.6
HMP-PS (with full connected graph)	48.6	41.4	54.1	72.4	72.8	79.5	84.3	60.8	42.4	58.2	40.1	48.9	58.6
HMP-PS (with prior graph)	53.5	**46.2**	53.5	75.7	74.4	**82.7**	86.0	61.7	49.8	59.8	43.6	49.0	61.3
HMP-PS (with gradient-based graph)	52.8	**46.2**	54.0	75.7	75.0	82.6	**86.4**	62.2	49.0	59.7	46.8	50.5	61.7
HMP-PS	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**

Table 4. AU correlations from anatomy defined in FCAS[10].

AU correlation	AUs
positive	(1,2), (4,7), (4,9), (6,12), (9,17), (15,24), (17,24), (23,24)
negative	(2,6), (2,7), (12,15), (12,17)

the highest averaged F1 score, $i.e.$, 63.4, and outperforms other state-of-the-art approaches. SRERL and MLCR exploited the semantic dependencies among AUs to enhance the deep neural networks. Both SRERL and LP-Net utilized additional landmark labels. Our HMP-PS still achieves better results, which indicates that our method is more effective to characterize the sophisticated dependencies among AUs. Both JAA-Net and ARL employed attention mechanism to find the regions of interest and we also utilize attention mechanism to select useful messages, which also validates the effectiveness of our algorithm to dynamically combine different messages. Our method performs well on hard AUs, like AU1, AU2 and AU23, which have lower F1 scores compared with other AUs. All the results on BP4D can validate the effectiveness of the proposed HMP-PS.

Table 2 provides the results of HMP-PS and the state-of-the-art methods on DISFA. HMP-PS also achieves best performance with averaged F1 score, $i.e.$, 61.0, which outperforms much better than SRERL by 5.1. All these results show the effectiveness of our method. The proposed P-MCMC sampling can generate multiple effective graph structures, which tend to have high performance. We consider more possible graph structures to represent dependencies among AUs and hybrid message passing provides more effective way to propagate the complementary information.

4.3. Ablation Study

To investigate the effectiveness of each part in our HMP-PS, we provide ablation study for further analysis.

Multiple Sampled Graph Structures: In our HMP-PS model, as shown in Figure 4, multiple hybrid message passing layers with different graph structures are combined to capture the dependencies among AUs. To validate the effectiveness of different graph structures generated by proposed P-MCMC sampling, we provide the results of ablation experiments with different types of graph in Table 3. First, we provide the results of baseline methods, $i.e$, HMP-PS(w/o graph) and HMP-PS (with full connected graph). Further, we use the prior knowledge in Table 4 defined in FACS [10] to construct the graph structure of our model, $i.e.$, HMP-PS (with prior graph). We can see the graph structure will directly affect the performance of our model on AU detection. Particularly, we provide the result of gradient-based method, $i.e$, HMP-PS (with gradient-based graph). The Gumbel-Sigmoid function [2] is employed to generate the discrete graph structure. Finally, we provide the result of HMP-PS. The performance of HMP-PS achieves significant improvement compared with other baseline methods, which validates the effectiveness of our P-MCMC sampling method to generate the graph structures and the advantage of multiple graph structures combination.

Table 5. The F1 score (in %) of ablation experiments on BP4D and DISFA reported by HMP-PS with different types of dynamic weights, $i.e.$, τ and p.

Methods	BP4D	DISFA
HMP-PS (w/o both weights)	60.8	59.0
HMP-PS (w/o dynamic weights τ)	62.5	60.4
HMP-PS (w/o dynamic weights p)	61.0	59.7
HMP-PS	**63.4**	**61.0**

Dynamic Weights for Multiple Graph Structures: In our HMP-PS model, multiple hybrid message passing networks with different graph structures are combined by dynamic weights τ. To validate the effectiveness of dynamic weights τ, we compare our HMP-PS with HMP-PS (w/o dynamic weights τ) in Table 5 on both BP4D and DISFA. The dynamic weights τ are helpful to improve the performance of our model and provide an effective way to combine the node features generated from different graph structures.

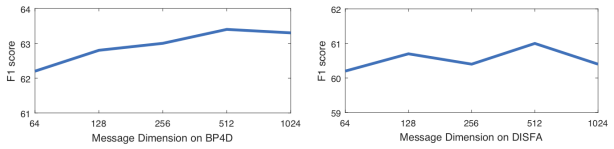
Dynamic Weights for Different Messages: For our HMP-PS, we dynamically combine different types of messages by using dynamic weights p. We also compare HMP-PS with HMP-PS (w/o dynamic weights p) in Table 5. The results indicate that dynamic weights p can significantly improve the performance of our model on AU detection. Besides, we also employ HMP-PS without both τ and p as a baseline method, which validates the effectiveness of τ and p.

Hybrid Messages: In our HMP-PS model, we employ three types of general approaches to estimate the messages, $i.e.$, summation message, differential message and concatenation message. To estimate the effectiveness of our hybrid message passing, we show the results with only one type of message respectively, $i.e.$, HMP-PS (with $h_i + h_j$), HMP-PS (with $h_i - h_j$) and HMP-PS (with cat$[h_i, h_j]$), as the

Table 6. The F1 score (in %) of ablation experiments on BP4D and DISFA reported by HMP-PS with different types of messages.

Methods	BP4D	DISFA
HMP-PS (with $h_i + h_j$)	61.6	58.4
HMP-PS (with $h_i - h_j$)	51.5	46.1
HMP-PS (with cat$[h_i, h_j]$)	62.3	58.9
HMP-PS (w/o $h_i - h_j$)	62.7	60.4
HMP-PS (hybrid features)	**63.4**	**61.0**

baseline methods in Table 6. Our HMP-PS achieves better performance than that with only one type of message, which indicates that combining different types of messages to propagate the hybrid information is helpful for the inference. Especially, the differential feature has the lowest F1 score on both BP4D and DISFA. We remove the differential feature from the HMP-PS, *i.e.*, HMP-PS (w/o $h_i - h_j$). HMP-PS outperforms HMP-PS (w/o $h_i - h_j$), which indicates that the differential message also contains useful information. The combination of different types of messages are necessary to improve the generalization capabilities of message passing neural networks.



(a) (b)

Figure 5. (a): The F1 scores of HMP-PS with different message dimensions on BP4D. (b): The F1 scores of HMP-PS with different message dimensions on DISFA.

4.4. Influence of Message Dimension

To investigate the influence of message dimension, we show the F1 scores with different message dimensions in Figure 5. Overall, different message dimensions will generate small influence on the result. However, proper message dimension is still helpful to achieve better performance. On both BP4D and DISFA, we achieve the best results when message dimension is set to 512. Small message dimension or large message dimension will have lower F1 scores.

Table 7. Training time for one epoch and inference time for all test data on BP4D.

Methods	Training	Inference
MLCR[23]	531s	77s
HMP-PS	685s	31s

4.5. Complexity analysis

The MCMC smapling is time-consuming, but can be done off-line. Once MCMC is done, the training time and inference time are not time consuming. In Table 7, we show training and inference time of HMP-PS and LP-Net on BP4D with a GeForce RTX 2080ti GPU. Our training time is slightly longer than a graph-based method, *i.e.*, MLCR but the inference is much faster.

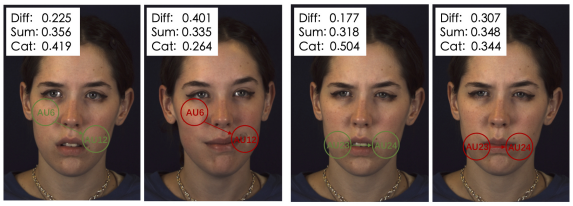


Figure 6. The visualized examples about the changes of dynamic weights p in terms of different AU occurrences. The left two images show the weights of the link from AU6 to AU12. The right two images show the weights of the link from AU23 to AU24. 'Diff' denotes the dynamic weight of differential message. 'Sum' denotes the dynamic weight of summation message. 'Cat' denotes the dynamic weight of concatenation message. The AUs with red color mean that these AUs occur and the AUs with green color mean that these AUs don't occur.

4.6. Visualization Analysis

To show the process that the hybrid message passing dynamically selects different messages, we visualize some examples in Figure 6. By comparing the first and the second images, we can see that the differential message will have larger weight if AU6 and AU12 occur. If both AU6 and AU12 don't occur, the concatenation message will have larger weight. As shown in the third and the fourth images, we have the similar conclusion. The weight for the differential messages is 0.177 when AU23 and AU24 don't appear. If AU23 and AU24 appear, the weights for the differential message will increase to 0.307 and the weights for concatenation message will decrease from 0.505 to 0.344. These examples indicate that our hybrid message passing neural network will adaptively select different messages based on different input data, which provide an effective way to characterize the dependencies among AUs.

5. Conclusion

In this paper, we focus on the graph structure learning and message propagation. First, we propose performance-driven MCMC sampling to generate multiple graph structures with high performance on AU detection. Second, we propose the hybrid message passing so as to propagate more complementary information. And then, we propose a novel framework (HMP-PS) for AU detection, which exploits more possible dynamic dependencies and achieves the state-of-the-art performance on two widely used AU detection databases. In the future, more effective ways inspired by belief propagation to extract the message can be applied to message passing neural networks.

Acknowledgments: We thank the support of China Scholarship Council. This work is supported in part by the National Natural Science Foundation of China under Grants U2003207 and 61921004. This work is supported in part by the U.S. National Science Foundation award IIS 1539012.

References

[1] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pages 4502–4510, 2016. 3, 4

[2] Jianlong Chang, Xinbang Zhang, Yiwen Guo, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Differentiable architecture search with ensemble gumbel-softmax. *arXiv preprint arXiv:1905.01786*, 2019. 2, 7

[3] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 3

[4] 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. 6

[5] 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, 2020. 2

[6] Zijun Cui, Yong Zhang, and Qiang Ji. Label error correction and generation through label relationships. In *AAAI*, pages 3693–3700, 2020. 2

[7] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016. 1, 3

[8] 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. 6

[9] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015. 1, 3, 4

[10] Rosenberg Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997. 1, 2, 7

[11] Yingruo Fan, Jacqueline CK Lam, and Victor On Kwok Li. Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution. In *AAAI*, pages 12701–12708, 2020. 1

[12] Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. *arXiv preprint arXiv:1903.11960*, 2019. 2

[13] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212*, 2017. 1, 2, 3, 4

[14] Paolo Giudici and Robert Castelo. Improving markov chain monte carlo model search for data mining. *Machine learning*, 50(1-2):127–158, 2003. 2, 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 5, 6

[16] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 2

[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*, volume 33, pages 8594–8601, 2019. 1, 2, 6

[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. 2, 6

[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

[20] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015. 3

[21] 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. 5

[22] Ignavier Ng, Zhuangyan Fang, Shengyu Zhu, and Zhitang Chen. Masked gradient-based causal structure learning. *arXiv preprint arXiv:1910.08527*, 2019. 2

[23] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Multi-label co-regularization for semi-supervised facial action unit recognition. In *Advances in Neural Information Processing Systems*, pages 909–919, 2019. 1, 2, 5, 6, 8

[24] 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. 6

[25] Nishant Sankaran, Deen Dayal Mohan, Srirangaraj Setlur, Venugopal Govindaraju, and Dennis Fedorishin. Representation learning through cross-modality supervision. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. 6

[26] 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. 1, 6

[27] 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. 1, 2, 6

[28] 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

[29] 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*, 2021. 2

[30] Tengfei Song, Suyuan Liu, Wenming Zheng, Yuan Zong, and Zhen Cui. Instance-adaptive graph for eeg emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2701–2708, 2020. 2

[31] Tengfei Song, Suyuan Liu, Wenming Zheng, Yuan Zong, Zhen Cui, Yang Li, and Xiaoyan Zhou. Variational instance-adaptive graph for eeg emotion recognition. *IEEE Transactions on Affective Computing*, (01):1–1, 2021. 2

[32] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 2018. 2

[33] Bowen Tang, Skyler T Kramer, Meijuan Fang, Yingkun Qiu, Zhen Wu, and Dong Xu. A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *Journal of Cheminformatics*, 12(1):1–9, 2020. 2

[34] Martin Trapp, Robert Peharz, Hong Ge, Franz Pernkopf, and Zoubin Ghahramani. Bayesian learning of sum-product networks. In *Advances in Neural Information Processing Systems*, pages 6347–6358, 2019. 3

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 5

[36] Robert Walecki, Vladimir Pavlovic, Björn Schuller, Maja Pantic, et al. Deep structured learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2017. 2

[37] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 5

[38] Shangfei Wang, Longfei Hao, and Qiang Ji. Knowledge-augmented multimodal deep regression bayesian networks for emotion video tagging. *IEEE Transactions on Multimedia*, 22(4):1084–1097, 2019. 1

[39] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. 2

[40] Sirui Xie, Hehui Zheng, Chunxiao Liu, and Liang Lin. Snas: stochastic neural architecture search. *arXiv preprint arXiv:1812.09926*, 2018. 2

[41] Yao Xue, Nilanjan Ray, Judith Hugh, and Gilbert Bigras. Cell counting by regression using convolutional neural network. In *European Conference on Computer Vision*, pages 274–290. Springer, 2016. 1

[42] Li Zhang, Dan Xu, Anurag Arnab, and Philip HS Torr. Dynamic graph message passing networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3726–3735, 2020. 3

[43] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013. 5

[44] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2314–2323, 2018. 1

[45] 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, 6