Dynamic Probabilistic Graph Convolution for Facial Action Unit Intensity Estimation

Tengfei Song[1], Zijun Cui[2], Yuru Wang[3*], Wenming Zheng[1*], and Qiang Ji[2]
[1]Southeast University
[2]Rensselaer Polytechnic Institute
[3]Northeast Normal University
songtf@seu.edu.cn, cuiz3@rpi.edu, wangyr915@nenu.edu.cn,
wenming_zheng@seu.edu.cn, qji@ecse.rpi.edu

Abstract

Deep learning methods have been widely applied to automatic facial action unit (AU) intensity estimation and achieved the state-of-the-art performance. These methods, however, are mostly appearance-based and fail to exploit the underlying structural information among AUs. In this paper, we propose a novel dynamic probabilistic graph convolution (DPG) model to simultaneously exploit AU appearances, AU dynamics, and their semantic structural dependencies for AU intensity estimation. Firstly, we propose to use Bayesian Network to capture the inherent dependencies among AUs. Secondly, we introduce probabilistic graph convolution that allows to perform graph convolution on the distribution of Bayesian Network structure to extract AU structural features. Finally, we introduce a dynamic deep model based on LSTM to simultaneously combine AU appearance features, AU dynamic features, and AU structural features for AU intensity estimation. In experiments, our method achieves comparable and even better performance with the state-of-the-art methods on two benchmark facial AU intensity estimation databases, i.e., FERA 2015 and DISFA.

1. Introduction

Facial muscle movements contain rich information related to human emotions, which are significant for human communication. Ekman and Friesen proposed a Facial Action Coding System (FACS) [8] to depict the movement of these facial muscles. FACS defines rules to annotate the intensities of different action units (AUs) and the AU intensities are quantified into 6 discrete levels.

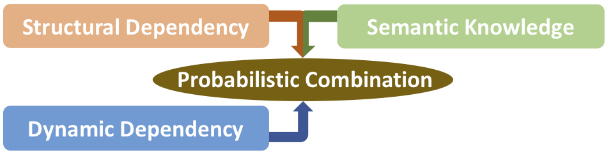


Figure 1. The illustration of combining the dynamic and structural dependencies with semantic knowledge for AU intensity estimation.

With the availability of high performance computing and the large scale datasets, deep neural network has been a dominant method in many computer vision tasks [35, 13]. The most commonly used deep neural networks, *i.e.*, Convolutional Neural Network (CNN) [19] and Long Short-Term Memory Network (LSTM) [15], show excellent performance in extracting discriminative local features and capturing the temporal dependencies, respectively. Their applications in the field of human behavior analysis also achieve much improvement, including human action recognition [26], expression recognition [41] and facial AU recognition [24, 36, 6, 21, 30]. However, their success highly rely on the large amount of training data, which is difficult to obtain for facial AU intensity estimation. The reason is that the process of facial AU intensity annotation is time-consuming and requires strong domain expertise. Moreover, the data distribution of AU intensity is generally imbalanced. Therefore, the performance of deep method on AU intensity estimation is limited under insufficient data. Rather than increasing training data for AU intensity estimation, the prior knowledge provides more generic information that is helpful to improve the performance of deep model. Because of the underlying facial anatomy and the need to form a coherent facial expression, a certain group of AUs may be activated under a specific expression and, at the same time, they may suppress the activity of other AUs [45]. For instance, cheek raiser and lip corner puller occur simultaneously for smile. Besides, the changes of AU over a short temporal period are usually observed as continuous

[*] Corresponding author

and smooth. Therefore, not only the AUs spatial dependencies, but also their temporal consistency, $i.e.$, dynamics, should be leveraged in estimating AU intensity.

Considering the aforementioned properties, we propose a novel dynamic probabilistic graph convolution (DPG) model that leverages semantic dependencies among AUs and integrates these semantic probabilistic information into a dynamic deep model for AU intensity estimation. As shown in Figure 1, semantic AU knowledge and structural dependency are combined within a unified probabilistic and dynamic frame. Firstly, we capture AU semantic dependencies with a Bayesian Network, which is advantageous in capturing the causal relationships and provides interpretable AU dependencies. Instead of using a deterministic graph, a sampling method is employed to estimate the posterior distribution of graph structure given the AU intensity estimation. We then introduce the proposed probabilistic graph convolution by incorporating all possible semantic structures with their probabilities into the conventional graph convolution. Finally, the probabilistic graph convolution model is embedded in a dynamic deep model, which simultaneously captures AU appearance, semantic structural and dynamic features for AU intensity estimation.

Our contributions can be summarized as follows:

- We propose a dynamic probabilistic deep framework to simultaneously integrate the AU appearance information, their dynamics and their structural dependencies for AU intensity estimation.
- We introduce a probabilistic graph convolution that allows learning structural features over probabilistic graphs.
- Our method achieves comparable and even better performance than the state-of-the-art methods on two benchmark AU intensity estimation datasets.

2. Related Works

2.1. Deep Model for AU intensity estimation

In recent years, deep neural network has been a popular method and shown great progress on AU intensity estimation due to its powerful representation ability [54, 22]. Walecki $et\ al.$ [42] integrated conditional random field (CRF) with CNN model to encode AU intensity dependencies, which demonstrated that the AU spatial relationship plays a crucial role in performance improvement. In addition to spatial relationships, some works also tried to exploit temporal consistency during the onset/offset process. Zhang $et\ al.$ [48] proposed to extract features with CNN and enforce temporal intensity order to improve the performance. Intuitively, temporal consistency will show more potential information in AU intensity estimation than detection. Therefore, more works exploited both spatial and temporal information in deep model for AU related tasks. Both Chu $et\ al.$ [3] and Zhang $et\ al.$ [49] developed deep networks, which extracted spatial representations by CNN, and modeled the temporal dependencies with LSTMs. The spatial relationship is obtained by the fully connected layer of CNN model, which is just concatenation of features from each AU. Sanchez $et\ al.$ [32] jointly perform AU localisation and intensity estimation via heatmap regression, which shows the stability against alignment errors. For AU intensity problem, integrating both spatial relationship and temporal consistency in a deep model based framework is expected to show performance improvement. Therefore, in this paper, we explore an effective deep graph structure to simultaneously characterize the structural and dynamic information of AUs.

2.2. Semantic Knowledge Model for AU

Some of the existing works encode the semantic knowledge to represent structural information and then enforce constraint to the learning procedure. Nicolle $et\ al.$ [29] exploited the underlying common structure between multiple tasks for AU intensity estimation. Eleftheriadis $et\ al.$ [10] proposed topological and global relational constraints on Multi-conditional Latent Variable Model, which exploit the structure-discovery capabilities of generative models. Zhao $et\ al.$ [52] derived two types of AU relations, $i.e.$, positive correlation and negative competition, under the statistic analysis of data. The pre-defined semantic knowledge can improve the performance of the multi-label classifier. Waleck $et\ al.$ [43] proposed to model the co-occurrence of AU intensity levels with the statistical framework of copula functions, which effectively captures correlations between facial features and co-occurrences of AUs. According to the prior knowledge from FACS [8], Li $et\ al.$ [20] proposed to learn a semantic graph from data so as to construct a knowledge graph coding AU correlation. Zhang $et\ al.$ [50] and Dong $et\ al.$ [47] integrated human knowledge to the existing model, which all demonstrate effectiveness of the semantic knowledge for AUs. Wang $et\ al.$ [44] proposed a method to obtain semantic knowledge from expression so as to enhance the dependencies among AUs. Integrating the knowledge model and the deep neural network in a unified framework is expected to learn more consistent representation such that Benitez $et\ al.$ [1] proposed a loss function combining the recognition of isolated and groups of AUs. In this paper, we integrate the semantic information into a deep model so as to characterize meaningful structural and dynamic dependencies AUs.

2.3. Graph Convolution

Graph convolution has been applied to many tasks, $i.e.$, image classification [7], action recognition [46] and object tracking [5]. Based on spectral graph theory [4, 39, 37, 38], graph convolution provides an advantageous representation for intrinsic relationships among different nodes. To avoid expensive calculation, Defferrard *et al.* [7] proposed an effective method to conduct graph convolution with low cost. The Chebyshev polynomial is employed to construct the graph convolution kernel. Shi *et al.* [33] proposed an adaptive method to generate the graph connections from data to capture spatial information. The model with the learned structure outperforms the model with the pre-defined structure. Jiang *et al.* [17] proposed the graph convolution based on Gaussian distribution, which characterized local variations of graph. More studies leverage graph convolution to exploit the relations among AUs. Niu *et al.* [30] employed AU relations via graph convolution to exploit more informative information from unlabeled face images. Fan *et al.* [11] proposed a dynamic graph convolution method to extract more discriminative features for AU intensity estimation. The existing methods to construct graph connections are deterministic model and mostly based on the natural structure of nodes. However, the deterministic graph cannot capture the underlying uncertain information in data. Linh *et al.* [25] and Eleftheriadi *et al.* [9] proposed the deep variational framework for the latent representation of AUs, which demonstrates the effectiveness of the probabilistic model for AU intensity estimation. Therefore, we introduce graph convolution in a probabilistic way to capture the graph's uncertain AU relations.

3. Proposed Method

In this section, we introduce the proposed dynamic probabilistic graph convolution (DPG) model that combines the dynamic discriminative deep model with probabilistic semantic knowledge related to the dependencies among AUs.

3.1. Probabilistic Graph Convolution From Semantic Knowledge

The proposed probabilistic graph convolution (PGC) is defined as the expected graph convolution over the posterior distribution of graph, which represents the distribution of semantic graph among AUs. Given AU intensity annotations $\mathcal{D}$ in training dataset, we employ Bayesian Network (BN) $\mathcal{G}$ to represent and encode probabilistic dependencies among AUs, *i.e.*, the posterior distribution $p(\mathcal{G}|\mathcal{D})$ of BN $\mathcal{G}$ via Markov Chain Monte Carlo (MCMC) sampling method. Since the correlations among AUs are important information to estimate the AU intensities, we then convert directed graph $\mathcal{G}$ to the undirected graph via moral graph and represent undirected graph by symmetric adjacency matrix A with the corresponding probability $p(A|\mathcal{D})$ for graph convolution.

3.1.1 Bayesian Network for AU Intensity Dependencies Encoding

We propose to employ a Bayesian Network (BN) to capture and encode the prior knowledge on AU relations as BN has strong interpretability in characterizing the relationships among AUs and it has been widely used in the past [44, 45, 23] for encoding AU dependencies. The Bayesian network, *i.e.*, the directed acyclic graph (DAG), can be defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes nodes and $\mathcal{E}$ for edges. Here, the nodes $\mathcal{V}$ represent AUs and the edges $\mathcal{E}$ represent the probabilistic dependencies among AUs.

Given AU intensity annotations $\mathcal{D}$, the posterior distribution $p(\mathcal{G}|\mathcal{D})$ of graph $\mathcal{G}$ captures probabilistic relationships among AUs. Directly computing $p(\mathcal{G}|\mathcal{D})$ is intractable and there exists no analytic solution. We use Markov Chain Monte Carlo (MCMC) sampling method [27] based on Metropolis Hasting algorithm (MH) [2], which generates N graph samples $\{\mathcal{G}_n\}_{n=1}^N$ to approximate $p(\mathcal{G}|\mathcal{D})$. The posterior $p(\mathcal{G}|\mathcal{D})$ can be written as $p(\mathcal{G}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{G})p(\mathcal{G})}{p(\mathcal{D})}$. Assuming uniform $p(\mathcal{G})$, $p(\mathcal{G}|\mathcal{D})$ is proportional to $p(\mathcal{D}|\mathcal{G})$, *i.e.*, $p(\mathcal{G}|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{G})P(\mathcal{G}) \propto p(\mathcal{D}|\mathcal{G})$.

The marginal likelihood $p(\mathcal{D}|\mathcal{G})$ can be written as

$$p(\mathcal{D}|\mathcal{G};\boldsymbol{\alpha}) = \int p(\mathcal{D}|\boldsymbol{\theta},\mathcal{G})p(\boldsymbol{\theta};\boldsymbol{\alpha})d\boldsymbol{\theta}, \tag{1}$$

where the parameters $\boldsymbol{\theta}$ of BN are used to represent the conditional probability distribution of each node given its parents. $\boldsymbol{\alpha}$ is the hyper-parameter for the Dirichlet distribution that represents the prior distribution of the parameters $\boldsymbol{\theta}$ and $\mathcal{D}$ contains AU intensity annotations. $\boldsymbol{\alpha}$ is usually set to be $\mathbf{1}$ for uniform distribution. The marginal likelihood can be solved analytically as

$$p(\mathcal{D}|\mathcal{G};\boldsymbol{\alpha}) = \prod_{i=1}^{N}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(N'_{ij}+\alpha_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N'_{ijk}+\alpha_{ijk})}{\Gamma(\alpha_{ijk})}, \tag{2}$$

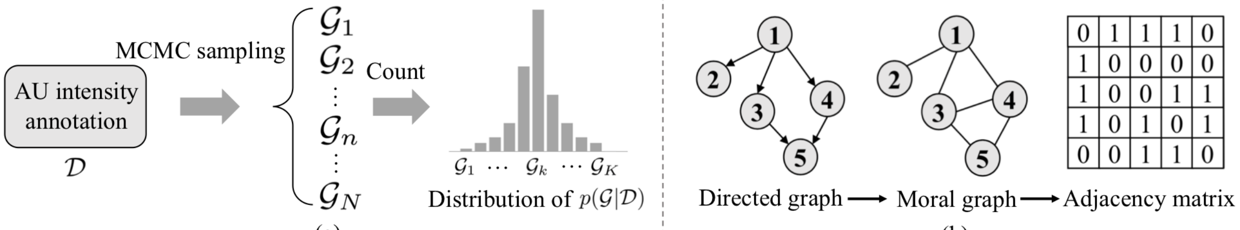
where i indicates the ith node in $\mathcal{G}$, j indicates the jth configuration of node i's parents, k indicates the kth configuration of node i, N'_{ijk} is the corresponding observed count and $\Gamma(\cdot)$ is Gamma function.

During the MH sampling, we generate a sequence of samples, following a Markov chain as determined the transition probability in Eq. (3). In particular, given a sample $\mathcal{G}_n$, a new sample $\mathcal{G}_{n+1}$ is proposed with the proposal probability

$$p^{pro}(\mathcal{G}_{n+1}|\mathcal{G}_n) = \frac{1}{|\mathcal{N}(\mathcal{G}_n)|} \tag{3}$$

if $\mathcal{G}_{n+1} \in \mathcal{N}(\mathcal{G}_n)$, and $p^{pro}(\mathcal{G}_{n+1}|\mathcal{G}_n) = 0$ if $\mathcal{G}_{n+1} \notin \mathcal{N}(\mathcal{G}_n)$. $\mathcal{N}(\mathcal{G}_n)$ denotes all neighboring DAGs of $\mathcal{G}_n$. Since $p(\mathcal{G}|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{G})$, we obtain $\frac{p(\mathcal{D}|\mathcal{G}_{n+1})}{p(\mathcal{D}|\mathcal{G}_n)} = \frac{p(\mathcal{G}_{n+1}|\mathcal{D})}{p(\mathcal{G}_n|\mathcal{D})}$ and the proposed DAG $\mathcal{G}_{n+1}$ is accepted with the acceptance probability

$$p^{acc}(\mathcal{G}_{n+1}|\mathcal{G}_n) = \min\{1, \frac{p(\mathcal{D}|\mathcal{G}_{n+1})}{p(\mathcal{D}|\mathcal{G}_n)}\frac{|\mathcal{N}(\mathcal{G}_n)|}{|\mathcal{N}(\mathcal{G}_{n+1})|}\} \tag{4}$$



0	1	1	1	0
1	0	0	0	0
1	0	0	1	1
1	0	1	0	1
0	0	1	1	0

(a) (b)

Figure 2. The illustration about (a) the generation of $p(\mathcal{G}|\mathcal{D})$ from AU annotation $\mathcal{D}$; (b) the generation of adjacency matrix from a directed graph to moral graph.

where we calculate likelihoods $p(\mathcal{D}|\mathcal{G}_n)$ and $p(\mathcal{D}|\mathcal{G}_{n+1})$ with Eq.(2) with $\boldsymbol{\alpha} = \mathbf{1}$. The Markov chain is left unchanged, *i.e.*, $\mathcal{G}_{n+1} := \mathcal{G}_n$ if the new graph $\mathcal{G}_{n+1}$ is not accepted, which means we accept the new samples based on the distribution of $p(\mathcal{G}|\mathcal{D})$. More details of MCMC sampling have been presented in supplemental material.

After a sufficiently large number of Burn-in samples, we can begin to collect N samples $\{\mathcal{G}_n\}_{n=1}^N$ to estimate the posterior distribution $p(\mathcal{G}|\mathcal{D})$ as shown in Figure 2 (a). The dependencies between AUs are significant to analyze the generation of AUs and some AUs have strong correlations. Thus, we employed the undirected graph to capture the mutual dependencies among AUs. To this goal, for each sampled DAG $\mathcal{G}_n$, we apply the moral graph [16] to obtain its corresponding undirected graph $\bar{\mathcal{G}}_n$. The moral graph is an undirected graph containing the same variables as the corresponding DAG and it contains the same independencies as the DAG. As shown in Figure 2 (b), the moral graph of a DAG is constructed by adding edges between all pairs of non-adjacent nodes having a common child and removing directions of all edges, *i.e.*, yielding the corresponding undirected moral graph. Based on which, we obtain the symmetric adjacency matrix A. After converting $\{\mathcal{G}_n\}_{n=1}^N$ to $\{\bar{\mathcal{G}}_n\}_{n=1}^N$, we compute the corresponding N symmetric adjacency matrices $\{A_n\}_{n=1}^N$. We use $A_{n,ij}$ denotes the element of the ith row and the jth column of A_n. $A_{n,ij} = 1$ means the ith node and the jth node are connected. Otherwise, $A_{n,ij}$ will be 0 if the ith node and the jth node are not connected. $\{A_n\}_{n=1}^N$ are representative for the underlying posterior distribution of the adjacency matrix $p(A|\mathcal{D})$.

3.1.2 Probabilistic Graph Convolution

Graph convolution provides a deep graphic representation to capture the underlying structure of data and gives a meaningful representation for the semantic relationships. The structural information allows to aggregate nodes with meaningful topological structure so as to explore more discriminative deep features for classification and regression tasks. Hammond *et al.* [12] conducted graph convolution in spectral domain, however, the calculation of graph Fourier transform is expensive. Defferrard [7] proposed to employ the K^{th}-order Chebyshev polynomial to approximate the computation of graph Fourier transformation. Recently, Kipf et al. [18] further leverage a linear approximation of graph convolution by reducing K to 1.

Let $H \in \mathbb{R}^{M\times d}$ denote the input signals, where each row of H denotes the feature vector for a node (AU), M is the number of nodes, d is the length of the feature vector of each node. One graph convolution layer g with input signal H can be defined as follows

$$H' = g(A, H) = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}HW), \tag{5}$$

where σ denote the activation function, $\tilde{A} = A + I$ is the adjacency matrix that contains the self-connection for each node, H' is the output, I is the identity matrix and $\tilde{D}$ is a diagonal degree matrix of $\tilde{A}$ with $\tilde{D}(i,i) = \sum_j \tilde{A}(i,j)$.

Conventional graph convolution is deterministic by nature. It applies only to a given fixed graph. For this research, instead of having a fixed graph, we obtain a distribution of graph, *i.e.*, $p(\mathcal{G}|\mathcal{D})$ also represented by $p(A|\mathcal{D})$ from the training data. The conventional deterministic graph convolution is therefore not applicable. To overcome this limitation, we introduce the probabilistic graph convolution:

$$\begin{aligned} H'_p &= \int p(A|\mathcal{D})\sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}HW)\, dA \\ &= E_{p(A|\mathcal{D})}(g(A, H)), \end{aligned} \tag{6}$$

where $g(A, H)$ represents the conventional convolution graph filter for an adjacency matrix A as defined in Eq. (5). The probabilistic convolution filter hence is defined as integration of the $g(A, H)$ over the posterior distribution of $p(A|\mathcal{D})$, *i.e.*, the expected $g(A, H)$ over $p(A|\mathcal{D})$, $E_{p(A|\mathcal{D})}(g(A, H))$. Integration of the graph structure distribution is intractable. Eq. (6) can be approximated by summation over the structure samples, *i.e.*,

$$H'_p \approx \frac{1}{N}\sum_{n=1}^{N} g(A_n, H) = \sum_{k=1}^{K} p(A_k|\mathcal{D})g(A_k, H), \tag{7}$$

where N is the number of samples we generated using MH method from Section 3.1.1. K is the number of unique graph structures. For a certain unique graph A_k, the probability of A_k given the data can be calculated by $\frac{m_k}{N}$, in which m_k is the frequency that the unique graph A_k appears

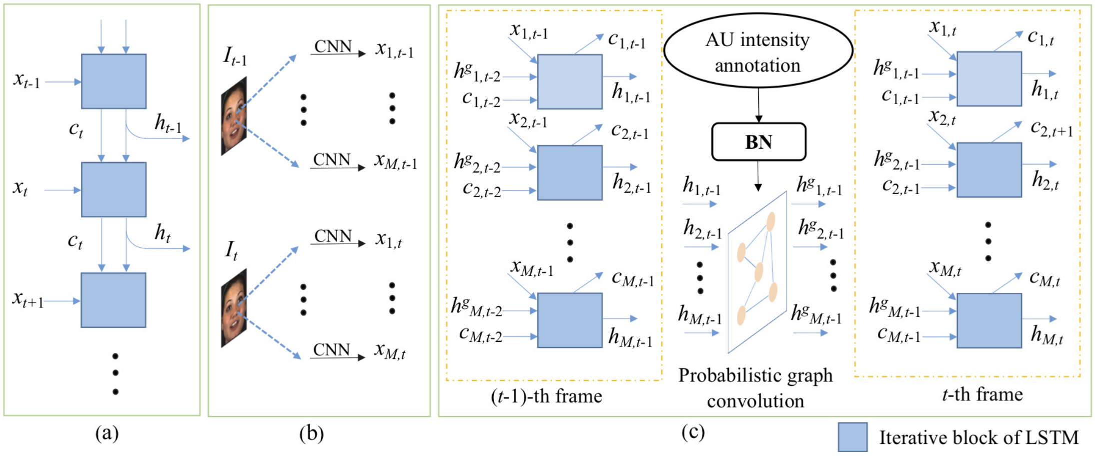


Figure 3. (a) One stream LSTM. (b) The framework to extract AU features. (c) The illustration of the dynamic probabilistic graph convolution model.

in N samples. The probabilistic graph convolution provides a novel way to capture the uncertain semantic dependencies among different nodes, which can be generalized to many different tasks.

3.2. Dynamic Probabilistic Graph Convolution Model

The changes among AUs are not independent and their spatial and temporal changing patterns are significant for AU intensity estimation. We incorporate the probabilistic graph convolution that encodes the semantic structural dependencies of AUs into multiple LSTMs that capture the long-term dependencies and short-term dependencies in the temporal dimension. The traditional one stream LSTM showed in Figure 3 (a) focuses on capturing the temporal dependencies of sequence. We devise a method to employ multiple LSTMs so as to further exploit their spatial structural relationships.

The Model In Figure 3 (b), given a continuous facial sequence $\mathcal{I} = \{I_1, I_2, ..., I_T\}$, we extract the feature vectors and each feature vector corresponds to one specific AU so as to obtain more accurate and independent features related to specific AUs. We employ ResNet50 [14] as the backbone network to extract the CNN features for each AU, represented by $\{x_{1,t}, x_{2,t}, ..., x_{M,t}\} \in \mathbb{R}^{1\times d}$, where d is the dimension of feature vector. For each AU, we pretrain a ResNet50 by using the training data.

To explore the spatial structural dependencies among AUs, multiple LSTMs are employed for AU intensity estimation and each LSTM focuses on the feature extraction of one type of AU. In the (t-1)th iterative process, the mth LSTM will generate a hidden state $h_{m,t-1}$, which contains the information of the mth AU in the (t-1)th image. In particular, we integrate the proposed probabilistic graph convolution into the LSTMs module such that this model can iteratively capture the spatial semantic dependencies among AUs. Let $H_{t-1} = [h_{1,t-1}; h_{2,t-1}; ...; h_{M,t-1}] \in \mathbb{R}^{M\times d}$ be the input hidden states for the tth iterative process. In the tth iterative step, we can model the spatial dependencies among AUs as following equation:

$$H^g_{t-1} = \sum_{k=1}^{K} p(A_k|\mathcal{D})\sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}_k\tilde{D}^{-\frac{1}{2}}HW), \tag{8}$$

where K is the number of unique graph structures, $\sum_{k=1}^{K} p(A_k|\mathcal{D})\sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}_k\tilde{D}^{-\frac{1}{2}}HW)$ denotes the proposed probabilistic graph convolution operation, $H^g_{t-1} = [h^g_{1,t-1}; h^g_{2,t-1}; ...; h^g_{M,t-1}] \in \mathbb{R}^{M\times d}$ is the output hidden states that fusing the semantic structural information, σ is the sigmoid function and $W_k \in \mathbb{R}^{d\times d}$ is signal transformation matrix. Since H^g_{t-1} has obtained the structural information among AUs of the (t-1)th frame, these information can propagate to the processing of the tth frame.

Simultaneously, in the tth iterative process, we want to capture the long-term and short-term temporal dependencies among AUs. The tth iterative process in temporal dimension can be expressed as

$$\begin{aligned}
&i_{m,t}=\sigma(w_{mxi}\cdot x_{m,t}+w_{mhi}\cdot h^g_{m,t-1}+w_{mci}\circ c_{t-1}+b_{mi})\\
&f_{m,t}=\sigma(w_{mxf}\cdot x_{m,t}+w_{mhf}\cdot h^g_{m,t-1}+w_{mcf}\circ c_{t-1}+b_{mf})\\
&u_{m,t}=tanh(w_{mxc}\cdot x_{m,t}+w_{m,hc}\cdot h^g_{m,t-1}+b_{mc})\\
&c_{m,t}=f_{m,t}\circ c_{m,t-1}+i_t\circ u_{m,t}\\
&o_{m,t}=\sigma(w_{mxo}\cdot x_t+w_{mho}\cdot h^g_{m,t-1}+w_{mco}\circ c_{m,t}+b_{mo})\\
&h_{m,t}=o_{m,t}\circ tanh(c_{m,t}), m=1,2,...,M.
\end{aligned} \tag{9}$$

$i_{m,t}$, $f_{m,t}$, $o_{m,t}$, $c_{m,t}$ denote the input gate, the forget gate, the output gate, the memory cell in the tth step of the mth LSTM, respectively. w_{mxi}, w_{mxf}, w_{mxc}, w_{mxo} are weight matrices specified for visual AU features. w_{mhi}, w_{mhf}, w_{mhc}, w_{mho} are weight matrices specified for spatial information among AUs. w_{mci}, w_{mcf}, w_{mco} denote the weight matrices for temporal dependencies. b_{mi}, b_{mf}, b_{mc}, b_{mo} are biases. $\circ$ denotes a point-wise product.

Specifically, $h^g_{m,t-1}$ contains the semantic structural information and the short-term temporal information, and c contains the long-term memory information. Therefore, this structure combining probabilistic graph convolution and multiple LSTMs can effectively capture the structural features and dynamic features simultaneously. The illustration of the dynamic probabilistic graph model is shown in Figure 3 (c). In particular, multiple LSTMs allow to extract more discriminative features from accurate locations of AUs respectively. Compared with the multiple LSTMs without information interaction, this structure extends the field of view of each LSTM such that more information can be fused to the feature extraction of next iterative process.

3.3. The Loss Function for AU Intensity Estimation

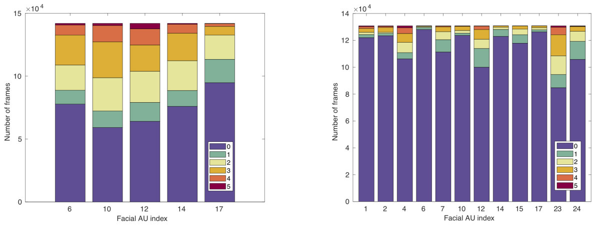
For each frame, the output hidden states, *i.e.*, $h_{m,t}$ of iterative blocks are connected with fully connected layers to reduce the dimensions and predict the intensity of AUs. The loss function can be defined as

$$loss = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{m=M}\|y_{m,t,n} - \bar{y}_{m,t,n}\|_2^2\right), \qquad (10)$$

in which $\|\cdot\|_2$ is the $l2$ norm, N is the number of training samples in one batch and $y_{m,t,n}$, $\bar{y}_{m,t,n}$ denote the true AU intensity and the predicted AU intensity respectively.

4. Experiments

4.1. Setting



(a) (b)

Figure 4. The distribution of facial AU intensity. (a) BP4D. (b) DISFA.

Datasets We evaluate the proposed method on two benchmark databases, *i.e.*, **FERA 2015**[40] and **DISFA** [28] database. FERA 2015 consists of 328 videos from 41 subjects when they are performing 8 tasks. There are about 140,000 frames, which are annotated with the intensity of 5 AUs. We employ the official Training/Development splits in our experiment, *i.e.*, 21 subjects for Training split and 20 subjects for Development split. DISFA is a spontaneous expression database, which contains 27 videos from 27 subjects when they are watching emotion elicitation videos. There are around 130,000 frames annotated with the intensity of 12 AUs. We employ 3-fold subject independent cross validation for evaluation with 18 subjects for training and 9 subjects for testing. Both FERA 2015 and DISFA contain 6 different AU intensity status from 0 to 5.

Data preprocessing For each AU, we train a Resnet50 [14] given the training data so as to get disentangled feature for a specific AU. We use the output of the final global pooling layer as the appearance feature of the specific AU. All the facial images are cropped and reshaped into 256x256. During training the ResNet50, we randomly crop the 256×256 images into 224×224, which are fed into ResNet50. The ResNet50 were pre-trained using training data. During training our DPG model, the parameters of ResNet50 will not be updated.

Evaluation metrics and hyperparameters We use Intra-Class Correlation (ICC(3,1) [34]) and Mean Absolute Error (MAE) as evaluation metrics. For our DPG model, we employ 5 LSTMs in FERA 2015 and 12 LSTMs in DISFA. The dimensions of the hidden states, memory cells of LSTMs are set to 256. The learning rates are set to 0.001 in FEAR 2015 and DISFA. During training and testing, the length of sequence is set to T=8 and the AUs intensities of each frame are predicted. For probabilistic graph convolution, we employed the top-5 possible unique graph structures and their normalized probabilities. The AU intensity annotations of training data are used to learn BN.

Comparison model First, we compare our DPG with the state-of-the-art methods, *i.e.*, **CCNN-IT** [42], **OR-CNN** [31], **OSVR** [53], **KJRE** [50], **BORMIR** [51], **KBSS** [48] and **CFLF** [49] , **2DC** [25] and **SCC** [11], for AU intensity estimation. Then, we perform the ablation study to compare the proposed **DPG** to some baseline methods. **Resnet50** [14] is our backbone network and **DPG-T** denotes baisc multiple LSTM model that captures temporal information without graph. Further, we present the proposed model using the graph convolution with the unique graph structure, which has the maximum posterior probability in $p(A|\mathcal{D})$ (**DPG-MPG**). Besides, we perform the graph convolution (**DPG-PK**) with the prior knowledge graph structure defined by the AU correlations in FACS [8] (Details in supplemental material). Both **DPG-MPG** and **DPG-PK** are deterministic models.

4.2. Results

4.2.1 Comparison with the state-of-the-art methods.

In Table 1, we compare the proposed method with the state-of-the-art deep methods. CCNN-IT, OR-CNN and KBSS are all deep methods that leverage structural or dynamic information. OSVR, KJRE, BORMIR and KBSS combine prior knowledge or semantic information for facial AU intensity estimation. 2DC also combines the deep model and probabilistic model. SCC is a deep graph method for AU intensity estimation. On FERA2015, our method achieves the best MAE on average and the same ICC on average with SCC. On DISFA, with more AU categories, our method

Table 1. Comparison to the state-of-the-art AU intensity estimation methods. Numbers in bracket and bold denote the best performance; numbers in bold denote the second best.

Database		FERA 2015						DISFA												
AU		6	10	12	14	17	Avg	1	2	4	5	6	9	12	15	17	20	25	26	Avg
ICC(3,1)	OSVR[53]	.65	.58	.78	.27	.45	.54	.21	.04	.25	.15	.23	.15	.31	.12	.07	.09	.62	.09	.19
	OR-CNN[31]	.74	.70	.85	.34	.51	.63	.01	.02	.21	.10	.47	.30	.76	.14	.21	.07	.84	.59	.31
	CCNN-IT[42]	.75	.69	.86	.40	.45	.63	.20	.12	.46	.08	.48	.44	.73	**.29**	**[.45]**	**[.21]**	.60	.46	.38
	KJRE[50]	.71	.61	**.87**	.39	.42	.60	.27	.35	.25	.33	.51	.31	.67	.14	.17	.20	.74	.25	.35
	BORMIR[51]	.73	.68	.86	37	.47	.62	.20	.25	.30	.17	.39	.18	.58	.16	.23	.09	.71	.17	.35
	KBSS[48]	**.76**	.75	.85	.49	.51	.67	.23	.11	.48	.25	.50	.25	.71	.22	.25	.06	.83	.41	.36
	CFLF[49]	**.76**	.70	.83	.41	**.60**	.66	.26	.19	.46	**.35**	.52	.36	.71	.18	.34	**[.21]**	.81	.51	.41
	2DC[25]	**.76**	.71	.85	.45	.53	.66	**.70**	**[.55]**	.69	.05	**.59**	**[.57]**	**[.88]**	**[.32]**	.10	.08	.90	.50	**.50**
	SCC[11]	.74	**[.82]**	.86	**[.68]**	.51	**[.72]**	**[.73]**	.44	**.74**	.06	.27	.51	.71	.04	.37	.04	**.94**	**[.78]**	.47
	DPG	**[.80]**	**.77**	**[.89]**	**.50**	**[.61]**	**[.72]**	.46	**.46**	**[.75]**	**[.63]**	**[.61]**	**.48**	**.84**	**.29**	**.44**	.18	**[.95]**	**.63**	**[.56]**
MAE	OSVR[53]	1.02	1.13	.95	1.35	.93	1.08	1.65	1.87	2.94	1.38	1.56	1.69	1.64	1.10	1.61	1.37	1.33	1.79	1.66
	OR-CNN[31]	**.56**	.72	.49	.95	.69	.68	.48	.45	.95	.04	.28	.23	**.27**	**.12**	.47	.12	.40	**[.32]**	.34
	CCNN-IT[42]	1.17	1.43	.97	1.65	1.08	1.26	.73	.72	1.03	.21	.72	.51	.72	.43	.50	.44	1.16	.79	.66
	KJRE[50]	.82	.95	.64	1.08	.85	.87	1.02	.92	1.86	.70	.79	.87	.77	.60	.80	.72	.96	.94	.91
	BORMIR[51]	.85	.90	.68	1.05	.79	.85	.88	.78	1.24	.59	.77	.78	.76	.56	.72	.63	.90	.88	.79
	KBSS[48]	**.56**	.65	**.48**	.98	.63	.66	.48	.49	.57	.08	**.26**	.22	.33	.15	.44	.22	.43	.36	.33
	CFLF[49]	.62	.83	.62	1.00	.63	.74	.33	.28	.61	.126	.35	.28	.42	.18	.29	.16	.53	.40	.33
	2DC[25]	.75	1.02	.66	1.44	.88	.95	.32	.39	.53	.26	.43	.30	**[.25]**	.27	.61	.18	.37	.55	.37
	SCC[11]	.61	**[.56]**	.52	**[.73]**	**[.50]**	**.58**	**[.16]**	**[.16]**	**[.27]**	**[.03]**	**[.25]**	**[.13]**	.32	.15	**[.20]**	**[.09]**	**.30**	**[.32]**	**[.20]**
	DPG	**[.50]**	**.61**	**[.43]**	**.81**	**.52**	**[.57]**	**.29**	**.26**	**.39**	**[.03]**	.27	**.14**	**.27**	**[.10]**	**.25**	**.11**	**[.24]**	.34	**.22**

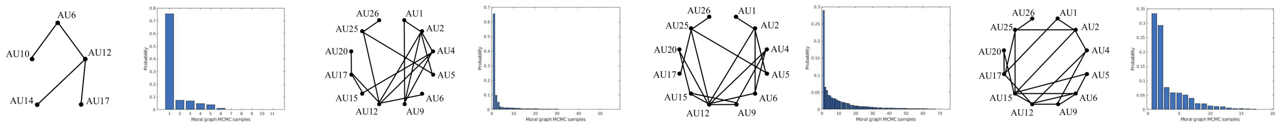
is advantageous to model their dependencies achieves the best performance in ICC on average. Especially, on both databases, our method outperforms over CCNN-IT, which is also a method combining semantic information with deep neural networks. Compared with KBSS that integrates dynamic information, our method achieves better results on both FERA2015 and DISFA. On DISFA, most of the AU intensity states are 0. SCC tends to correctly recognize more AUs with 0 state and our method tend to recognize more AUs with high intensities. This is why our method have higher ICC score than SCC but SCC has a lower MAE score. Especially, our DPG achieves much better ICC score, 0.56 than 2DC, *i.e.*, 0.5, which is the best result among comparison methods on DISFA. CFLF is also a method that takes into consideration of spatial relationships among AUs and our method outperforms CFLF on both FERA2015 and DISFA. Compared with FERA2015, training data on DISFA is more unbalanced and the results demonstrate our method is more powerful to deal with these unbalanced data by integrating semantic structural information with dynamic information. The aforementioned results further validate the effectiveness of our method.

4.2.2 Ablation study

The results for ablation study are presented in Table 2. The proposed DPG model achieves the best average performance on both FERA2015 and DISFA databases. To evaluate the dynamic information, DPG-T (with dynamic information) is compared with ResNet50 (without dynamic information) and DPG-T achieves better performance than ResNet50. Especially, we conduct our model with different types of graph convolution. The model with probabilistic graph convolution (DPG) has shown improvement over the model with a deterministic prior knowledge graph (DPG-PK) and the model with the maximum posterior probability graph (DPG-MPG). The probabilistic graph convolution (DPG) considers more uncertain graph structures such that richer structural information can be fused into deep model, which demonstrates the effectiveness of the probabilistic graph. In summary, DPG provides an effective deep architecture to characterize the temporal dynamic information and the spatial structural information simultaneously. Besides, DPG also integrates the semantic knowledge, *i.e.*, the AU dependencies, into deep architecture so as to capture more comprehensive information.

Table 2. The ICC and MAE results using different baseline methods on FERA2015 and DISFA for facial AU intensity estimation.

Database		FERA 2015						DISFA												
AU		6	10	12	14	17	Avg	1	2	4	5	6	9	12	15	17	20	25	26	Avg
ICC(3,1)	Resnet50	.76	.73	.88	.42	.59	.68	.43	.39	.63	.62	.53	.36	.82	**.33**	.41	**.18**	.92	.50	.51
	DPG-T	.77	.76	.86	.48	.58	.69	.45	.44	.70	.32	**.61**	.47	.82	.25	.44	.17	.92	.61	.52
	DPG-MPG	.79	.75	.87	.49	.59	.70	**.46**	.45	.74	.40	.60	.46	.83	.27	**.45**	**.18**	.93	.62	.53
	DPG-PK	.78	.76	.87	.49	.58	.70	**.46**	.45	.73	.51	**.61**	.47	.83	.27	.44	**.18**	.93	.62	.54
	DPG	**.80**	**.77**	**.89**	**.50**	**.61**	**.72**	**.46**	**.46**	**.75**	**.63**	**.61**	**.48**	**.84**	.29	.44	**.18**	**.95**	**.63**	**.56**
MAE	Resnet50	.62	.70	.45	.92	.55	.64	.38	.38	.65	.07	.36	.34	.35	.19	.38	.21	.38	.50	.35
	DPG-T	.55	.65	.50	.85	.57	.62	.32	.31	.48	.06	.30	.19	.32	.13	.29	.13	.34	.38	.27
	DPG-MPG	**.50**	.62	.44	**.80**	**.52**	.58	.31	.30	.44	.06	.29	.17	.31	.13	.29	.12	.32	**.33**	.26
	DPG-PK	**.50**	.62	.48	**.80**	.53	.59	.32	.30	.44	.05	.29	.18	.31	.13	.28	.12	.31	.38	.26
	DPG	**.50**	**.61**	**.43**	.81	**.52**	**.57**	**.29**	**.26**	**.39**	**.03**	**.27**	**.14**	**.27**	**.10**	**.25**	**.11**	**.24**	.34	**.22**



(a) (b) (c) (d)

Figure 5. The graph with the maximum probability and the graph probability distributions. (a) FERA 2015; (b) The first fold of DISFA; (c) The second fold of DISFA; (d) The third fold of DISFA;

4.2.3 Computational complexity

The probabilistic graph convolution (PGC) is linear summation of traditional graph convolution (GCN) such that the time complexity will not increase too much. With a RTX2080 Ti GPU, we predict the result of 90 sequences (one batch) on Tensorflow. The time cost is 0.22s with GCN and 0.29s with PGC.

4.3. The Visualization of Graph

To further analyze the semantic dependencies among AUs, we visualize the moral graph with the maximum probability and show the probabilities of all unique graphs using DPG in Figure 5.

For FEAR 2015, as shown in Figure 5 (a), AU6 has connections with AU10 and AU12, which means that cheek raiser (AU6) has the strong connections with upper lip raiser (AU10) and lip corner puller (AU12). At the same time, lip corner puller (AU12) also has the strong connections with dimpler (AU14) and lower lip depressor (AU17). Aforementioned discussions are based on the graph with maximum probability and there are still many weak connections among AUs that are not displayed. Besides, some facial examples that reflect the AU dependencies are shown in supplemental material.

For DISFA, we present three different graphs and probability distributions in Figure 5 (b), (c) and (d) since we employ 3-fold subject independent cross validation to evaluate the proposed method. Some connections among AUs are different and the probability distributions also reflect differences, which means the distribution of AU dependencies will change in terms of different people. The proposed Bayesian-based method captures different semantic dependencies and can be generalized into different training data. Though we leverage the AU intensity annotation of training data of 3-fold experiments to generate graphs, there are still many common connections among AUs. For instance, AU1 is connected with AU2, AU2 is connected with AU12 and AU25 is connected with AU26. Besides, AU2 and AU12 have at least four connections with other AUs.

5. Conclusion

We proposed a novel DPG model to simultaneously capture AU appearance features, AU dynamic features, and AU semantic structural features for AU intensity estimation. By performing the Bayesian Network and deep graph neural networks, we offer promising new directions to combine the probabilistic model with the deep model. The distribution of the graph sampled from data is presented to exploit useful semantic knowledge and capture underlying uncertain dependencies among AUs. We hope our DPG model can inspire more studies on exploiting the underlying semantic knowledge for AU intensity estimation.

Acknowledgments: We thank the support of China Scholarship Council. This work is supported in part by the National Natural Science Foundation of China under Grant U2003207, and in part by the Jiangsu Frontier Technology Basic Research Project under the grant BK20192004. This work is supported in part by the U.S. National Science Foundation award CNS 1629856.

References

[1] Carlos Fabian Benitez-Quiroz, Yan Wang, and Aleix M Martinez. Recognition of action units in the wild with deep nets and a new global-local loss. In *ICCV*, pages 3990–3999, 2017. 2

[2] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995. 3

[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. IEEE, 2017. 2

[4] Fan RK Chung and Fan Chung Graham. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997. 2

[5] Zhen Cui, Youyi Cai, Wenming Zheng, Chunyan Xu, and Jian Yang. Spectral filter tracking. *IEEE Transactions on Image Processing*, 28(5):2479–2489, 2018. 2

[6] 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. 1

[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. 2, 3, 4

[8] Paul Ekman and Wallace V Friesen. *Manual for the facial action coding system*. Consulting Psychologists Press, 1978. 1, 2, 6

[9] Stefanos Eleftheriadis, Ognjen Rudovic, Marc Peter Deisenroth, and Maja Pantic. Variational gaussian process autoencoder for ordinal prediction of facial action units. In *Asian Conference on Computer Vision*, pages 154–170. Springer, 2016. 3

[10] Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3792–3800, 2015. 2

[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. 3, 6, 7

[12] David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011. 4

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[14] 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. 5, 6

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1

[16] Finn V Jensen et al. *An introduction to Bayesian networks*, volume 210. UCL press London, 1996. 4

[17] Jiatao Jiang, Zhen Cui, Chunyan Xu, and Jian Yang. Gaussian-induced convolution for graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4007–4014, 2019. 3

[18] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 4

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[20] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. *arXiv preprint arXiv:1904.09939*, 2019. 2

[21] 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

[22] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eacnet: 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

[23] Yongqiang Li, S Mohammad Mavadati, Mohammad H Mahoor, Yongping Zhao, and Qiang Ji. Measuring the intensity of spontaneous facial action units with dynamic bayesian network. *Pattern Recognition*, 48(11):3417–3427, 2015. 3

[24] 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 Conference on Computer Vision and Pattern Recognition*, pages 10924–10933, 2019. 1

[25] Dieu Linh Tran, Robert Walecki, Stefanos Eleftheriadis, Bjorn Schuller, Maja Pantic, et al. Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3190–3199, 2017. 3, 6, 7

[26] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016. 1

[27] David Madigan, Jeremy York, and Denis Allard. Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, pages 215–232, 1995. 3

[28] 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. 6

[29] Jeremie Nicolle, Kevin Bailly, and Mohamed Chetouani. Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. In *2015 11th IEEE International Conference and Workshops on Automatic Face*

and Gesture Recognition (FG), volume 6, pages 1–6. IEEE, 2015. 2

[30] 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, 3

[31] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016. 6, 7

[32] Enrique Sánchez-Lozano, Georgios Tzimiropoulos, and Michel Valstar. Joint action unit localisation and intensity estimation through heatmap regression. *arXiv preprint arXiv:1805.03487*, 2018. 2

[33] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2019. 3

[34] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979. 6

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[36] 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. 1

[37] 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

[38] 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

[39] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2018. 2

[40] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–8. IEEE, 2015. 6

[41] Raviteja Vemulapalli and Aseem Agarwala. A compact embedding for facial expression similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5683–5692, 2019. 1

[42] 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, 6, 7

[43] Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Copula ordinal regression for joint estimation of facial action unit intensity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4902–4910, 2016. 2

[44] Shangfei Wang, Longfei Hao, and Qiang Ji. Facial action unit recognition and intensity estimation enhanced through label dependencies. *IEEE Transactions on Image Processing*, 28(3):1428–1442, 2018. 2, 3

[45] Ziheng Wang, Yongqiang Li, Shangfei Wang, and Qiang Ji. Capturing global semantic relationships for facial action unit recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3304–3311, 2013. 1, 3

[46] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[47] 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*, pages 5108–5116, 2018. 2

[48] 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. 2, 6, 7

[49] Yong Zhang, Haiyong Jiang, Baoyuan Wu, Yanbo Fan, and Qiang Ji. Context-aware feature and label fusion for facial action unit intensity estimation with partially labeled data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 733–742, 2019. 2, 6, 7

[50] Yong Zhang, Baoyuan Wu, Weiming Dong, Zhifeng Li, Wei Liu, Bao-Gang Hu, and Qiang Ji. Joint representation and estimator learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3457–3466, 2019. 2, 6, 7

[51] Yong Zhang, Rui Zhao, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7034–7043, 2018. 6, 7

[52] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, 2015. 2

[53] Rui Zhao, Quan Gan, Shangfei Wang, and Qiang Ji. Facial expression intensity estimation using ordinal information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3466–3474, 2016. 6, 7

[54] Yuqian Zhou, Jimin Pi, and Bertram E Shi. Pose-independent facial action unit intensity regression based on multi-task deep transfer learning. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 872–877. IEEE, 2017. 2