# Uncertain Graph Neural Networks for Facial Action Unit Detection

**Tengfei Song**[1,2], **Lisha Chen**[3], **Wenming Zheng**[1], **Qiang Ji**[3]
[1]Key Laboratory of Child Development and Learning Science, Ministry of Education, Southeast University, Nanjing, China
[2]School of Information Science and Engineering, Southeast University, Nanjing, China
[3]Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, New York, USA
songtf@seu.edu.cn, chenl21@rpi.edu, wenming_zheng@seu.edu.cn, qji@ecse.rpi.edu

## Abstract

Capturing the dependencies among different facial action units (AU) is extremely important for the AU detection task. Many studies have employed graph-based deep learning methods to exploit the dependencies among AUs. However, the dependencies among AUs in real world data are often noisy and the uncertainty is essential to be taken into consideration. Rather than employing a deterministic mode, we propose an uncertain graph neural network (UGN) to learn the probabilistic mask that simultaneously captures both the individual dependencies among AUs and the uncertainties. Further, we propose an adaptive weighted loss function based on the epistemic uncertainties to adaptively vary the weights of the training samples during the training process to account for unbalanced data distributions among AUs. We also provide an insightful analysis on how the uncertainties are related to the performance of AU detection. Extensive experiments, conducted on two benchmark datasets, i.e., BP4D and DISFA, demonstrate our method achieves the state-of-the-art performance.

## Introduction

Analyzing and recognizing the facial expressions are significant to realize the natural human-robot interaction in the field of artificial intelligence. Facial Action Unites (AU) defined in the Facial Action Coding System (FACS) (Ekman 1997) characterize the facial muscle movements, which underpin multiple facial expressions. Thus, researchers increasingly conduct AU detection to estimate the human emotion and facial behaviors (Danelakis, Theoharis, and Pratikakis 2018; Wang, Hao, and Ji 2018).

Recently, deep neural networks (He et al. 2016; Simonyan and Zisserman 2014; Ertugrul, Le Yang, and Cohn 2019) have been the dominant methods to perform the computer vision tasks with large scale datasets and high performance computing. Many studies (Li, Abtahi, and Zhu 2017; Niu et al. 2019a) employed deep neural networks to extract appearance features for AU detection and the performance has been remarkably improved compared with traditional methods (Jiang, Valstar, and Pantic 2011; Simon et al. 2010). Most of the existing deep learning methods for AU detection are deterministic models, which rely on the large amount of training data with precise annotation. For AU detection, the annotation of the ground truth requires strong domain expertise and hence the annotations are often noisy (Zhang et al. 2013). Besides, the annotations are often unbalanced as some AUs appear frequently while other AUs rarely appear. All these make the AU detection still a challenging task.

Typically, AU detection can be treated as a multi-label classification task in terms of different AU types with some intrinsic dependencies among different AUs. Rather than designing more complex networks to learn features in a data-driven manner to detect each AU independently, recent studies (Zhang et al. 2018; Wang et al. 2013) begin to exploit the intrinsic dependencies among different facial muscles due to the underlying facial anatomy and the need to form a meaningful expression. For example, 'raise cheek' and 'pull lip corners' generally appear simultaneously when we show the happiness expression. On the other hand, 'raise cheek' and 'stretch mouth' are not likely to appear simultaneously. Recently, more researchers (Niu et al. 2019a,b; Corneanu, Madadi, and Escalera 2018; Wang, Gan, and Ji 2017) capture these dependencies among AUs with a graph, which provides an effective way to combine the appearance features with semantic dependencies. The graph-based method can capture the complex geometric and semantic relationships between nodes based on strong mathematical graph theory (Chung and Graham 1997). Most of these existing works about AU detection employ some prior knowledge to define a fixed graph, which cannot effectively capture the dynamic and individual dependencies among AUs. The graph can also be learned by an end-to-end manner (Shi et al. 2019), which can capture better graph dependencies from the distribution of the data. However, it still assumes the same graph structure for all images. Inspired by the attention mechanism, the graph attention network (Veličković et al. 2017) was proposed recently to focus on partial set of the most relevant edges based on input, which can adaptively provide larger weight for the edge with high contribution in the graph. Although these methods provide effective ways to exploit the individual dependencies among AUs and can extract more discriminative features, most of them are deterministic deep models for AU detection. Generally, the deterministic model cannot effectively capture the data uncertainty, including label noise, and cannot quantify its prediction confidence.

To overcome the limitation of deterministic model, we propose to capture the dependencies among AUs in a probabilistic manner. The probabilistic model can generally be applied to capture the underlying uncertainties in the data as well as in the model, which can provide more reliable results and is robust to the noise. Furthermore, we parameterize the adjacency matrix of graph convolution such that we can learn the intrinsic graph dependencies among AUs from the data. Inspired by attention mechanism, we generate the probabilistic mask to adaptively vary the importance of the dependencies among AUs for each input image. The mean and the variance of the probabilistic mask depend on the input such that the model can adjust the mean and measure the uncertainties based on the input. As a result, our probabilistic model captures both the aleatoric uncertainty of the data and the epistemic uncertainty of model. Finally, we can use the estimated uncertainty to generate an adaptive weighted loss function to vary the weights of training samples to effectively handle the data imbalance problem. Especially, we provide insight discussion about the relationships between epistemic uncertainties and the performance, which also validates the effectiveness of the proposed model.

The contributions in this paper can be summarized as follows:

- We propose an uncertain graph neural network that employs a probabilistic mask to simultaneously capture the importance of the dependencies among AUs for each input and estimate the prediction uncertainties.
- Based on the uncertainties, we propose an adaptive weighted loss function, which varies the weights for training samples to address the data imbalance problem.
- Our method achieves the state-of-the-art performance on two benchmark datasets, i.e., BP4D and DISFA.

## Related Works

In this part, we will review some existing related works, including the general deep models for AU detection, AU detection with AU relationships modeling and graph neural networks for AU detection.

### Deep Models for AU Detection

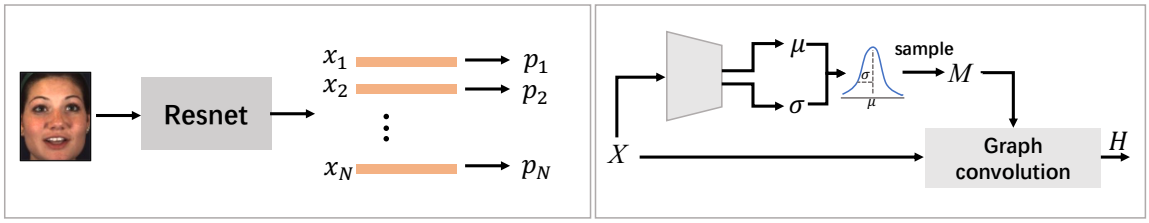
Recently, most of works about AU detection are based on deep models due to the powerful representation ability. (Gudi et al. 2015) applied the deep convolutional neural networks to AU detection, which provided the discriminative representation for spatial face features. (Shao et al. 2018) proposed attention-based deep model to adaptively select the regions with high contributions for joint AU detection and face alignment. A multi-scale structure was also proposed to represent the features from different levels to help AU detection. (Tang et al. 2017) proposed to employ multiple deep neural networks for multi-view facial AU detections. (Tu, Yang, and Hsu 2019) integrated the identity information into the training of deep neural network. The identity-label facial images are used to solve the identity-based intro-class variation of AU detection. (Li et al. 2019b) proposed a self-supervised deep method to detect the AU occurrence in videos. The muscle movement between two neighbor face images was served as the self-supervisory signals to learn the temporal representation. All these deep models are deterministic, which can provide discriminative representation. However, they don't take into consideration of the underlying uncertain information, which is significant for the real-world AU detection.

### AU Relationships Modeling

As observed in FACS (Ekman 1997), for specific expression, some AU activations are strongly correlated such that many previous works exploit the intrinsic dependencies among AUs for AU detection. (Zhao, Chu, and Zhang 2016) proposed to combine the region learning with multi-label learning in a unified deep network, which simultaneously captures the important region and the AU dependencies. (Cui, Zhang, and Ji 2020) exploited the relationships between AU labels via a Bayesian network to correct the noisy labels. The improvement of the results demonstrated the importance of AU relations. (Li et al. 2019a) proposed to learn an offline graph from data, combined with the prior knowledge from FACS so as to construct a knowledge graph coding the AU correlations. Integrating the knowledge model and the deep neural network in a unified framework is expected to learn more consistent representation. (Walecki et al. 2017) appended a conditional random field (CRF) model to the end of VGG model, and jointly learned the parameters. CRF was applied to encoding the AU dependencies with deep learning. (Wang, Ding, and Peng 2020) utilized the dependencies between AUs, the dependencies between expressions and AUs, and the dependencies between facial features and AUs for AU detection under nonfull annotation. Compared to offline knowledge, learning the AU representation and semantic knowledge together will obtain more consistent semantic representation. Rather than employing the common relationships, we propose to adaptively generate the relationships based on data, which can better reflect the individual differences and exploit the intrinsic dependencies among AUs.

### Graph Neural Network

Graph neural network (GNN) is a popular machine learning model based on the spectral theory (Chung and Graham 1997) that incorporates the dependencies between data points. Recently, GNN has been widely applied to many computer vision tasks, e.g., action detection (Yan, Xiong, and Lin 2018), image classification (Defferrard, Bresson, and Vandergheynst 2016) and objection tracking (Cui et al. 2018). Most existing works defined a constant graph structure based on prior information. (Niu et al. 2019a) employed graph convolution for AU detection with a constant probabilistic matrix calculated from AU labels. (Shi et al. 2019; Song et al. 2018, 2020) proposed to directly learn the graph during the end-to-end training. The learned graph can better capture the dependencies between nodes in a data-driven manner. (Fan, Lam, and Li 2020) integrated the co-occurrences of different AU into graph neural network, which dynamically computed the correspondences from feature maps of deep network. However, the constant graph




(a) (b)

Figure 1: (a): The framework to extract AU features; (b): The framework of uncertain graph convolution.

can't characterize the individual differences for AU detection. (Veličković et al. 2017) proposed graph attention networks (GAT) that strengthen the links with high contribution in graph, which provide a more flexible way to construct the graph. All these methods are based on deterministic graphs, which fails to reflect the underlying uncertain information. For AU detection, the labels are often noisy such that exploiting uncertain dependencies is significant to obtain reliable results. Recently, (Zhang et al. 2019; Vaida and Patil 2020) employed Bayesian methods to capture uncertainty in graph neural networks. In this paper, we build graph model in a probabilistic way to simultaneously capture the dependencies among AUs and the underlying uncertainties, where uncertainties are used to weight the loss function to improve the classification performance under imbalanced data.

## Method

In this section, we will introduce the uncertain graph neural network (UGN), which exploits the dependencies among AUs and the underlying uncertainties simultaneously. Specifically, we obtain the AU features via ResNet18 (He et al. 2016) and employ the proposed uncertain graph method to extract discriminative features for AU detection. Further, we propose an adaptive weighted loss based on the epistemic uncertainties to weight the training samples to alleviate the imbalance issue in the training data.

### The Generation of AU Features

The first step for AU detection is to extract corresponding effective deep appearance features for each AU. As shown in Figure 1 (a), we use ResNet18 (He et al. 2016) as the basic deep model to extract the AU features. We obtain $N$ feature vectors $X = \{x_1, ..., x_N\} \in \mathbb{R}^{d\times N}$ corresponding to $N$ AUs from ResNet18. $d$ is the dimension for each feature vector. Each feature vector is fed into a fully connected layer to predict the probability of one specific AU. Here, the predicted probability for the $i$-th AU can be written as

$$p_i = \text{sigmoid}(w_i^T x_i + b_i), \tag{1}$$

in which $w_i \in \mathbb{R}^{1\times d}$ and $b_i \in \mathbb{R}$ are parameters of $i$-th fully connected layer and a sigmoid function is applied to output the probability. For each AU, a binary cross-entropy loss is employed. And the total loss contains all the loss functions from each AU with a group of weights to balance the training data. The total loss function can be written as

$$\mathcal{L}_{au} = -\sum_{i=1}^{N} \tau_i [y_i' \log\ p_i + (1 - y_i') \log(1 - p_i)] \tag{2}$$

where $\tau_i$ is the weight for the $i$-th AU to balance the training data (Shao et al. 2018) and $y_i'$ is the ground truth for the $i$-th AU, with 1 denoting AU appears and 0 denoting AU does not appear.

### Definition of the Uncertain Graph

A graph can be defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, in which $\mathcal{V}$ denotes the nodes and $\mathcal{E}$ denotes the edges. Generally, the adjacency matrix $A \in \mathbb{R}^{N\times N}$ can be applied to represent the relationships among the nodes, i.e., the edges in a graph. $N$ is the number of nodes. (Defferrard, Bresson, and Vandergheynst 2016) constructed a constant graph based on the spatial distance between nodes and (Shi et al. 2019) proposed to learn the graph from data. To adaptively capture the important dependencies in a graph for each input, the graph attention network (GAT) (Veličković et al. 2017) was proposed by employing a mask to select the useful edges and depress the noisy edges. The graph convolution with the adaptive mask can be written as

$$H = M \odot (\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{\frac{1}{2}}) X^T W, \tag{3}$$

in which $\tilde{A} = A + I$ is the adjacency matrix that includes the self-connection for each nodes, $I \in \mathbb{R}^{N\times N}$ is the identity matrix, $H \in \mathbb{R}^{N\times d_{out}}$ is the output, $\odot$ is the element-wise production and $\tilde{D}$ is a diagonal degree matrix of $\tilde{A}$, with $\tilde{D}(i,j) = \sum_j \tilde{A}(i,j)$. $A$ is a symmetric weighted adjacency matrix to characterize the strength of the links and $W \in \mathbb{R}^{d\times d_{out}}$ is the parameter of neural networks. We also set $A$ to be the parameter that can be optimized during the training process. To obtain a non-negative symmetric adjacency matrix, $A$ is reparameterized by $A = ReLU(B+B^T)$, in which $B \in \mathbb{R}^{N\times N}$ is a parameter matrix. Here, $M \in \mathbb{R}^{N\times N}$ is a weighted mask characterizes the importance of the edges in a graph and $M$, i.e., $M(X)$, is dependent on the features of nodes and it hence varies with input. Rather than employing a deterministic model, we assume each element of $M$, i.e., $M_{ij}$, is a random variable such that we can capture the underlying uncertain information. $M_{i,j}$ can be reparameterized by

$$M_{i,j} = \text{sigmoid}(z_{i,j}), \tag{4}$$

in which $z_{i,j}$ is a random variable that follows Gaussian distribution such that the conditional probability distribution of $z_{i,j}$ can be written as

$$p(z_{i,j}|X) = \mathcal{N}(\mu_{i,j}(X), \sigma^2_{i,j}(X)). \tag{5}$$

$z_{i,j}$ is dependent on the input $X$ such that we employ a mapping function to calculate the mean $\mu_{i,j}(X)$ and the standard deviation $\sigma_{i,j}(X)$ of $z_{i,j}$ that can be represented as

$$\begin{aligned} \mu_{i,j}(X) &= a_\mu[Px_i||Px_j], \\ \sigma_{i,j}(X) &= \log(\exp(a_\sigma[Px_i||Px_j]) + 1), \end{aligned} \tag{6}$$

in which $P \in \mathbb{R}^{d' \times d}$ is a parameter matrix, $a_\mu \in \mathbb{R}^{1 \times 2d'}$ and $a_\sigma \in \mathbb{R}^{1 \times 2d'}$ are two parameter vectors and $[||]$ denotes the concatenation operation. Given the input graph features $X$, $\mu_{i,j}$, i.e., $\mu_{i,j}(X)$ is employed to generate the $M_{i,j}$, i.e., $M_{i,j}(X)$, which adaptively measures the importance between the $i$-th node and the $j$-th node in graph. $\sigma^2_{i,j}$ characterizes the uncertainty between the $i$-th node and the $j$-th node. The probabilistic mask $M$ can adaptively capture the importance of the links in a graph by the mean and characterize the uncertainties of the input data by the variance simultaneously.

## Learning the Uncertain Graph for AU Detection

For AU detection, we employ two graph convolutional layers with uncertain graph and a fully connected layer to predict AU occurrence. Since $M$ is probabilistic matrix reparameterized by latent variable matrix $z$, for the prediction, we employ the expected probabilities over the distribution of $z$ to predict AU occurrence. The AU labels can be predicted by

$$\begin{aligned} p(y_i|X) &= \mathbb{E}_{z \sim p(z|X)}[p(y_i|z, X)] \\ &= \int p(y_i|z, X)p(z|X)\, dz \\ &\approx \frac{1}{S}\sum_{s=1}^{S} f(A, M^s, X), \end{aligned} \tag{7}$$

in which $f(\cdot)$ denotes two uncertain graph convolution layers, one fully connected layer and one softmax. $S$ is the number of samples. $A$ is the learned adjacency matrix and $M^s = \text{sigmoid}(z^s)$ is the adaptive mask for the $s$-th sampling. Each element of $z$ follows Gaussian distribution in Eq. (5) and $z^s$ is the $s$-th sampled matrix from the random variable matrix $z$. $p(y_i|X)$ is the expected probability over $z \sim p(z|X)$ and the integral in Eq. (7) is intractable. We hence use sample average from $S$ samples to approximate the expectation.

During the training, we should optimize the expected loss function over the distribution $p(z|X)$. However, the sampling operation is not differentiable. We employed the reparameterization trick (Kingma and Welling 2013) such that we can calculate the gradients to obtain an optimal $p(z|X)$. Each element of $z$ can be represented as

$$z_{i,j} = \mu_{i,j}(X) + \sigma_{i,j}(X)\epsilon_{i,j},\ \epsilon_{i,j} \sim \mathcal{N}(0, 1), \tag{8}$$

in which $\epsilon_{i,j}$ is a random variable that follows standard normal distribution and $\epsilon_{i,j}$ is independent with the model parameters. Rather than directly generating a lot of samples to estimate the distribution of $\epsilon_{i,j}$ for each training process, we employ the stochastic method that samples one time for each $\epsilon_{i,j}$ to obtain the sampled $z$, i.e., $z^s$. Therefore, the total loss function for one image $X$ can be written as

$$\begin{aligned} \mathcal{L}_X = \sum_{i=1}^{N} \mathcal{L}_i = -\sum_{i=1}^{N} [& y'_i \log(p(y_i|z^s, X)) \\ &+ (1 - y'_i)\log(1 - p(y_i|z^s, X))], \end{aligned} \tag{9}$$

in which $\mathcal{L}_i$ denotes the loss of the $i$-th AU and $y'_i$ is the ground truth of the $i$-th AU.

## Adaptive Weighted Loss Function Based on the Uncertainties

The uncertainty can be divided into two types, i.e., aleatoric uncertainty and the epistemic uncertainty. Aleatoric uncertainty characterizes the uncertainties of intrinsic noise in the data and epistemic uncertainty characterizes the uncertainties of model. We can use our model to estimate these two types of uncertainties. The uncertainties, especially the epistemic uncertainties, are related to the model performance. As epistemic uncertainty measures data density in the training data, we employ a larger weight for the data with high epistemic uncertainties to automatically counter data imbalance.

The total uncertainties can be represented as the combination of aleatoric uncertainties and the epistemic uncertainties. We can quantify the epistemic uncertainties by evaluating the mutual information (Depeweg et al. 2017) between $y_i$ and $z$ for the $i$-th AU by

$$\begin{aligned} &\mathcal{I}[y_i, z|X] \\ &= \mathcal{H}[\mathbb{E}_{p(z|X)}[p(y_i|z, X)]] - \mathbb{E}_{p(z|X)}[\mathcal{H}[p(y_i|z, X)]] \\ &= \mathcal{H}[\frac{1}{S}\sum_{s=1}^{S} p(y_i|X, z^s)] - \frac{1}{S}\sum_{s=1}^{S} \mathcal{H}[p(y_i|X, z^s)], \end{aligned} \tag{10}$$

in which $\mathcal{H}[p] = -(p \log\, p + (1-p)\log\,(1-p))$ denotes the function to calculate the entropy, $\mathcal{H}[p(y_i|X)]$ is the total uncertainty, $\mathbb{E}_{p(z|X)}[\mathcal{H}[p(y_i|z, X)]]$ is the aleatoric uncertainty and $\mathcal{I}[y_i, z|X]$ is the epistemic uncertainty.

During the training process, we divide the training data into many batches and each batch contains $K$ training samples. Before training every batch of data, we will estimate the adaptive weights based on the epistemic uncertainties. The adaptive weights can be calculated by

$$w_k = 1 + \frac{e^{\sum_{i=1}^{N} \mathcal{I}[y_i, z|X^k]}}{\sum_{k=1}^{K} e^{\sum_{i=1}^{N} \mathcal{I}[y_i, z|X^k]}} \tag{11}$$

in which $w_k$ is the adaptive weight for the $k$-th training sample, $\sum_{i=1}^{N} \mathcal{I}[y_i, z|X^k]$ is the epistemic uncertainties for the $k$-th training sample, which is the summation of the epistemic uncertainties of different AUs. And then, we can use

| AU | F1-score | | | | | Accuracy | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ResNet18 | ResNet-BL | GN-D | UGN | UGN-B | ResNet18 | ResNet-BL | GN-D | UGN | UGN-B |
| 1 | 46.3 | 51.5 | 51.7 | 52.7 | **54.2** | 74.6 | **79.9** | 76.8 | 76.6 | 78.6 |
| 2 | 38.1 | 39.6 | 44.3 | 45.5 | **46.4** | 80.1 | 78.3 | 77.3 | 77.2 | **80.2** |
| 4 | 52.7 | 49.2 | 54.9 | 55.6 | **56.8** | 79.0 | 79.5 | 79.7 | 79.0 | **80.0** |
| 6 | 74.4 | 73.2 | **77.3** | 76.6 | 76.2 | 75.6 | 74.9 | **77.2** | 76.9 | 76.6 |
| 7 | 74.5 | 76.4 | 76.4 | **76.7** | **76.7** | 71.8 | **72.7** | 71.9 | 72.3 | 72.3 |
| 10 | 82.6 | **83.7** | 83.2 | 82.8 | 82.4 | 78.5 | **80.6** | 79.0 | 78.3 | 77.8 |
| 12 | 83.9 | 85.8 | 85.8 | 85.8 | **86.1** | 81.8 | 83.8 | 83.3 | 83.4 | **84.2** |
| 14 | 59.8 | 59.9 | **65.3** | 65.2 | 64.7 | 61.0 | 63.0 | 63.6 | **64.3** | 63.8 |
| 15 | 48.1 | 50.2 | 49.2 | 51.1 | **51.2** | 83.5 | **85.0** | 82.5 | 83.3 | 84.0 |
| 17 | 55.8 | 58.0 | 62.6 | **63.3** | 63.1 | 70.1 | 71.0 | 71.8 | 72.6 | **72.8** |
| 23 | 42.9 | 42.2 | 47.8 | 47.3 | **48.5** | 81.0 | 83.0 | **83.4** | 82.4 | 82.8 |
| 24 | 45.2 | 50.2 | 52.7 | 53.3 | **53.6** | 84.7 | 86.0 | 85.9 | 85.5 | **86.4** |
| Avg | 58.7 | 59.9 | 62.6 | 63.0 | **63.3** | 76.8 | 78.1 | 77.7 | 77.7 | **78.2** |

Table 1: The F1 score and the accuracy (in %) for the recognition of 12 AUs with different baseline methods on the BP4D data.

these adaptive weights to constrain the loss function to train the model. The total loss function will be

$$\mathcal{L} = \sum_{k=1}^{K} w_k \mathcal{L}_{X^k}, \tag{12}$$

in which $\mathcal{L}_{X^k}$ is the loss for the $k$-th training sample, i.e., $X^k$. The hard training sample will have higher epistemic uncertainty thus assigned a larger weight $w_k$.

# Experiment

## Experiment Setting

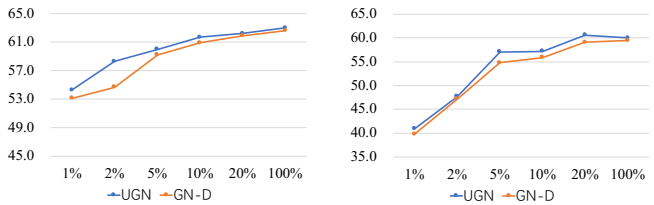
**Dataset** We evaluate our method on two widely used benchmark datasets for AU detection, i.e., BP4D (Zhang et al. 2013) and DISFA (Mavadati et al. 2013).

BP4D is a facial AU dataset that contains 41 participants with 23 females and 18 males. Each participant is involved in 8 sessions such that there are totally 328 videos for these 41 participants. There are about 140,000 frames with AU labels. 12 AUs are evaluated following the subject exclusive 3-fold cross validation, which is the same experiment protocol as (Shao et al. 2018) (Li et al. 2019a).

DISFA contains of 27 videos recorded from 15 males and 12 females and each subject has 4,845 images. For each image, AU intensities from 0 to 6 are annotated. Follow the same setting with the previous works (Shao et al. 2018) (Li et al. 2019a), the image with intensities equal or greater than 2 are considered as the occurrence of AU. 8 AUs are evaluated using subject exclusive 3-fold cross validation.

**Evaluation Metrics** We employed the F1 score and the accuracy to estimate our method. F1 score is generally used in binary classification, and involves the precision $p$ and the recall $r$. F1 score can be estimated by $F1 = 2\frac{p \cdot r}{p+r}$. For AU detection, the F1 score of the positive category is presented.

**Training Details** We train the ResNet18 model to obtain the AU features and ResNet18 is pre-trained on ImageNet (Deng et al. 2009). For each image in this experiment, we crop the face region and resize each cropped face into 256×256. During training to generate the feature vectors for AUs, the facial images are randomly cropped to 224×224 for ResNet and the random horizontal flipping is also utilized for data augmentation. The dimension of each AU feature, i.e., $d$, is 512. Given the extracted AU features, we build two uncertain graph layers to predict AU occurrence. The node dimension of the output of each uncertain graph layer, i.e., $d_{out}$ is 64. The learning rate of the uncertain graph model is set to 0.01 and the batch size for all experiments is set to 16. We implement the uncertain graph neural network with Tensorflow on a GeForce RTX 2080 GPU.




(a) (b)

Figure 2: The F1 scores (in %) of UGN and GN-D with 1%, 2%, 5%,10%, 20 % and 100% training data respectively on BP4D and DISFA. (a) The result on BP4D; (b) The result on DISFA.

## Comparison with Baseline Methods

In this section, we provide the performance of some baseline methods including the ResNet18, the ResNet18 with balanced loss (Resnet-BL), the deterministic graph attention network (GN-D), the proposed uncertain graph neural network (UGN) and the UGN with adaptive weighted loss (UGN-B).

Table 1 has shown the F1 scores and the accuracies on BP4D with the baseline methods and our proposed method. The ResNet18 with balanced weights for the loss, i.e., ResNet-BL, achieves better performance than ResNet in both F1 score and the accuracy on average, since these weights alleviate the influence of the unbalanced data in training dataset. Compared with ResNet-BL, the deterministic GN-D model has higher F1 score but lower averaged accuracy. This is because ResNet-BL tends to correctly recognize more AUs with 0 state and GN-D tends to correctly recognize more AUs with 1 state, i.e., AU occurrence. The

| AU | F1-score | | | | | | Accuracy | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ROI | JAA | ARL | SRERL | LP-Net | UGN-B | JAA | ARL | SRERL | UGN-B |
| 1 | 36.2 | 47.2 | 45.8 | 46.9 | 43.4 | **54.2** | 74.7 | 73.9 | 67.6 | **78.6** |
| 2 | 31.6 | 44.0 | 39.8 | 45.3 | 38.0 | **46.4** | **80.8** | 76.7 | 70.0 | 80.2 |
| 4 | 43.4 | 54.9 | 55.1 | 55.6 | 54.2 | **56.8** | 80.4 | **80.9** | 73.4 | 80.0 |
| 6 | 77.1 | **77.5** | 75.7 | 77.1 | 77.1 | 76.2 | **78.9** | 78.2 | 78.4 | 76.6 |
| 7 | 73.7 | 74.6 | 77.2 | **78.4** | 76.7 | 76.7 | 71.0 | 74.4 | **76.1** | 72.3 |
| 10 | **85.0** | 84.0 | 82.3 | 83.5 | 83.8 | 82.4 | **80.2** | 79.1 | 80.0 | 77.8 |
| 12 | 87.0 | 86.5 | 86.6 | **87.6** | 87.2 | 86.1 | 85.4 | 85.5 | **85.9** | 84.2 |
| 14 | 62.6 | 61.9 | 58.8 | 63.9 | 63.3 | **64.7** | **64.8** | 62.8 | 64.4 | 63.8 |
| 15 | 45.7 | 43.6 | 47.6 | **52.2** | 45.3 | 51.2 | 83.1 | **84.7** | 75.1 | 84.0 |
| 17 | 58.0 | 60.3 | 62.1 | **63.9** | 60.5 | 63.1 | 73.5 | **74.1** | 71.7 | 72.8 |
| 23 | 38.3 | 42.7 | 47.4 | 47.1 | 48.1 | **48.5** | 82.3 | **82.9** | 71.6 | 82.8 |
| 24 | 37.4 | 41.9 | **55.4** | 53.3 | 54.2 | 53.6 | 85.4 | 85.7 | 74.6 | **86.4** |
| Avg | 56.4 | 60.0 | 61.1 | 62.9 | 61.0 | **63.3** | **78.4** | 78.2 | 74.1 | 78.2 |

Table 2: The F1 score and the accuracy (in %) for the recognition of 12 AUs with the state-of-the-art methods on the BP4D dataset. Since ROI and LP-Net do not report the accuracy results, we just show their F1-score.

| AU | F1-score | | | | | | Accuracy | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ROI | JAA | ARL | SRERL | LP-Net | UGN-B | JAA | ARL | SRERL | UGN-B |
| 1 | 41.5 | 43.7 | **43.9** | 45.7 | 29.9 | 43.3 | 93.4 | 92.1 | 76.2 | **95.1** |
| 2 | 26.4 | 46.2 | 42.1 | 47.8 | 24.7 | **48.1** | **96.1** | 92.7 | 80.9 | 93.2 |
| 4 | 66.4 | 56.0 | 63.6 | 59.6 | **72.7** | 63.4 | 86.9 | **88.5** | 79.1 | **88.5** |
| 6 | **50.7** | 41.4 | 41.8 | 47.1 | 46.8 | 49.5 | 91.4 | 91.6 | 80.4 | **93.2** |
| 9 | 8.5 | 44.7 | 40.0 | 45.6 | **49.6** | 48.2 | 95.8 | 95.9 | 76.5 | **96.8** |
| 12 | **89.3** | 69.6 | 76.2 | 73.5 | 72.9 | 72.9 | 91.2 | **93.9** | 87.9 | 93.4 |
| 25 | 88.9 | 88.3 | **95.2** | 84.3 | 93.8 | 90.8 | 93.4 | **97.3** | 90.9 | 94.8 |
| 26 | 15.6 | 58.4 | 66.8 | 43.6 | **65.0** | 59.0 | 93.2 | **94.3** | 73.4 | 93.8 |
| Avg | 48.5 | 56.0 | 58.7 | 55.9 | 56.9 | **60.0** | 92.7 | 93.3 | 80.7 | **93.4** |

Table 3: The F1 score and the accuracy (in %) for the recognition of 8 AUs with the state-of-the-art methods on the DISFA dataset.

learned adjacency matrix can capture the intrinsic dependencies among AUs and the adaptive mask strengthens the dependencies with high contributions depresses the noisy dependencies based on different input data. Our proposed UGN achieves better performance than GN-D in both F1 score and the average accuracy. The probabilistic representation adaptively captures the AU dependencies and the underlying uncertainties, which are helpful to mitigate the noise labels and train a more robust model. With the adaptive weighted loss, UGN-B achieves the best result in both F1 score and the average accuracy. The adaptive weighted loss performs well on AU1, AU2 and AU23, which have more imbalanced data distribution than other AUs. The improvement of F1 score on imbalanced AUs demonstrates the effectiveness of the proposed adaptive weighted loss based on the epistemic uncertainties.

Specifically, to further evaluate the effectiveness of the propose probabilistic model under different amount of training data, we uniformly select partial training data to train the ResNet18 and employ graph-based methods to predict the AU occurrence. In Figure 2, we provide the F1 score using the proposed UGN and the GN-D with 1%, 2%, 5%, 10%, 20% and 100% training data respectively on BP4D and DISFA. Especially, the proposed UGN even achieves higher F1 score, i.e., 60.6 %, with 20 % training data than that with 100 % training data. Selecting proper ratio of training data may reduce the noisy data and our UGN can benefit from the selected data so as to extract more discriminative features. The proposed uncertain graph method outperforms the deterministic graph method, which demonstrates that the UGN has better generalization capability than deterministic graph model.

## Comparison with State-of-the-art

We compare our proposed uncertain graph method against the state-of-the-art methods for single-image based AU detection on BP4D and DISFA. These methods include ROI (Li, Abtahi, and Zhu 2017), JAA (Shao et al. 2018), ARL (Shao et al. 2019), SRERL (Li et al. 2019a) and LP-Net (Niu et al. 2019b). Specifically, all these methods are deep learning methods. To provide a fair comparison, we employ the results of ROI, JAA, ARL, SRERL and LP-Net reported in (Li, Abtahi, and Zhu 2017; Shao et al. 2018, 2019; Li et al. 2019a; Niu et al. 2019b)

Table 2 provides the F1 score and accuracy results of different state-of-the-art methods. Our method achieves the highest averaged F1 score and average accuracy on DISFA. On BP4D, our method achieves the highest F1 score and JAA has higher averaged accuracy than our method. This is because JAA correctly recognizes more AUs with 0 state and our method tends to correctly recognize more AUs with 1 state, i.e., AU occurrence. ROI and JAA are all deep models to adaptively select the region of interest for AU detection. Our model adaptively selects the important dependencies in

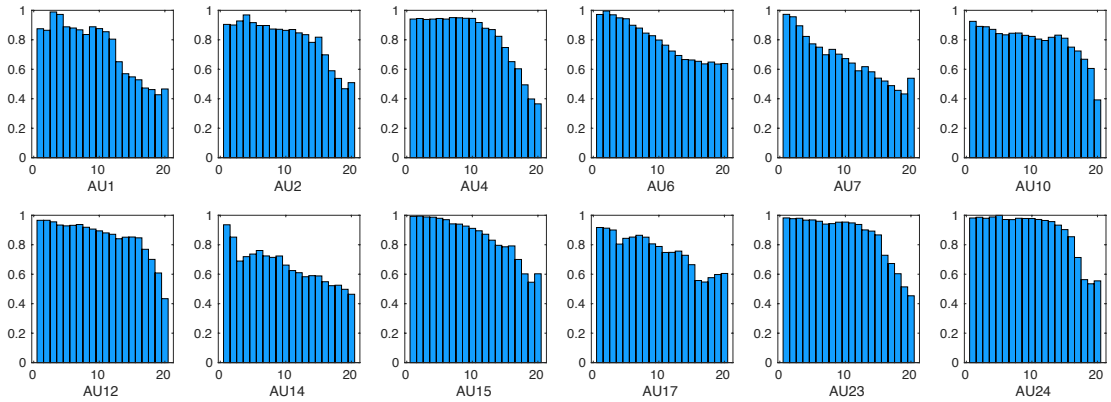



Figure 3: The histograms of the epistemic uncertainties with the accuracies for 12 AUs on BP4D. The horizontal axis represents the uncertainties from low to high and the vertical axis is the accuracy.

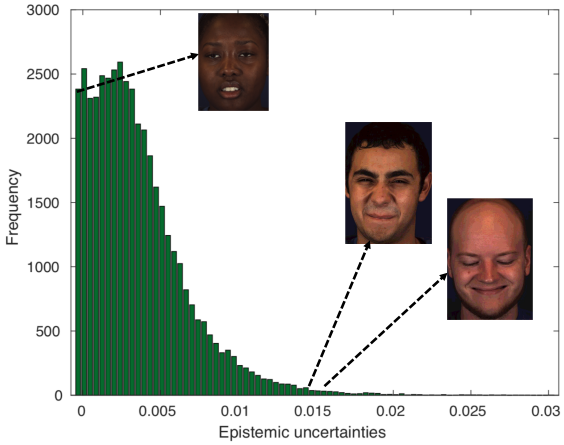



Figure 4: The epistemic uncertainty distribution for our UGN model on BP4D dataset.

a graph, which takes into consideration of the relationships among AU. ARL also takes advantage of the relation among AUs, and both SRERL and LP-Net integrate the semantic information to the graph convolution to capture the graph dependencies among AUs. Compared with these deterministic models to capture the relationships among AUs, our probabilistic uncertain model achieves better results since our model can characterize more uncertain information to mitigate the noise data and noise labels. Especially for the AUs with lower F1 scores, like AU1 and AU2, our method can work better. All these results demonstrate the effectiveness of our probabilistic model.

## Uncertainty Analysis

To further investigate the relationships between the epistemic uncertainties and the accuracies, we provide the histograms of the epistemic uncertainties with the accuracies for 12AUs on BP4D dataset, which have been shown in Figure 3. For each AU, we divided the testing data of the first fold into equal 20 bins from low uncertainties to high uncertainties, which means that each bin contains the same number of testing data. And then, we calculate the accuracy of each bin. We can see that the epistemic uncertainties of AUs are strongly related to the accuracies. For each AU, the data with high epistemic uncertainties tend to have low accuracies and the data with low epistemic uncertainties tends to have high accuracies such that it is reasonable to find the hard samples based on epistemic uncertainties so as to alleviate the influence of imbalance data. The adaptive weighted loss relies on the relations between the uncertainties and the accuracies.

In Figure 4, we present the epistemic uncertainty distribution for our UGN model on BP4D dataset. The horizontal axis of Figure 4 characterizes the epistemic uncertainties and vertical axis represents the frequency, i.e., data density. Further, we visualize some examples to figure out the difference between the facial images with high epistemic uncertainties and the facial images with low epistemic uncertainties. We can see clearly that the epistemic uncertainty negatively correlates with the data density. For the facial images with low epistemic uncertainties, most of AUs do not appear. But for the facial images with high epistemic uncertainties, there are rich expressions and multiple AUs are activated. Generally, for AU detection, the data is unbalanced and there are only a small percentage of positive samples. Typically, these positive samples are extremely important to train a robust model. Based on these relations, we provide larger weights for these images with high epistemic uncertainties, which helps improve the performance of our model.

## Conclusion

In this paper, we introduce probabilistic uncertain graph model and the adaptive weighted loss function to train the model. The probabilistic mask adapts the graph to individual images and captures the underlying uncertain information. Further, we use the epistemic uncertainties to select the hard training samples, which can alleviate the unbalanced data problem. Especially, for these AUs with lower F1 scores, like AU1 and AU2, our method works better. Experiments on benchmark datasets show the effectiveness of our model and we hope the proposed UGN model can inspire more researches on exploiting the uncertain information to more applications.

## Acknowledgments

We thank the support of China Scholarship Council. We also thank the resources provided by Rensselaer Polytechnic Institute. This work is supported in part by the National Key Research and Development Program of China under Grant 2018YFB1305200, and in part by the Jiangsu Frontier Technology Basic Research Project under the grant BK20192004. This work is supported in part by the National Science Foundation awards IIS 1539012 and CNS 1629856.

## Ethical Impact

The method in this paper can positively benefit various applications, including human computer interactions, medical diagnosis, animation and robotics. The datasets in this paper are made publicly available with a license that allows free usage for research purposes. Different from facial recognition, the risk of facial AU estimation is minimal.

## References

Chung, F. R.; and Graham, F. C. 1997. *Spectral graph theory*. 92. American Mathematical Soc.

Corneanu, C.; Madadi, M.; and Escalera, S. 2018. Deep structure inference network for facial action unit recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 298–313.

Cui, Z.; Cai, Y.; Zheng, W.; Xu, C.; and Yang, J. 2018. Spectral filter tracking. *IEEE Transactions on Image Processing* 28(5): 2479–2489.

Cui, Z.; Zhang, Y.; and Ji, Q. 2020. Label Error Correction and Generation through Label Relationships. In *AAAI*, 3693–3700.

Danelakis, A.; Theoharis, T.; and Pratikakis, I. 2018. Action unit detection in 3D facial videos with application in facial expression retrieval and recognition. *Multimedia Tools and Applications* 77(19): 24813–24841.

Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, 3844–3852.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Depeweg, S.; Hernández-Lobato, J. M.; Doshi-Velez, F.; and Udluft, S. 2017. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. *arXiv preprint arXiv:1710.07283* .

Ekman, R. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

Ertugrul, I. O.; Le Yang, L. A. J.; and Cohn, J. F. 2019. DPAttNet: Dynamic patch-attentive deep network for action unit detection. *Frontiers in computer science* 1.

Fan, Y.; Lam, J. C.; and Li, V. O. K. 2020. Facial Action Unit Intensity Estimation via Semantic Correspondence Learning with Dynamic Graph Convolution. In *AAAI*, 12701–12708.

Gudi, A.; Tasli, H. E.; Den Uyl, T. M.; and Maroulis, A. 2015. Deep learning based facs action unit occurrence and intensity estimation. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, 1–5. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiang, B.; Valstar, M. F.; and Pantic, M. 2011. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Face and Gesture 2011*, 314–321. IEEE.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

Li, G.; Zhu, X.; Zeng, Y.; Wang, Q.; and Lin, L. 2019a. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8594–8601.

Li, W.; Abtahi, F.; and Zhu, Z. 2017. 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*, 1841–1850.

Li, Y.; Zeng, J.; Shan, S.; and Chen, X. 2019b. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10924–10933.

Mavadati, S. M.; Mahoor, M. H.; Bartlett, K.; Trinh, P.; and Cohn, J. F. 2013. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing* 4(2): 151–160.

Niu, X.; Han, H.; Shan, S.; and Chen, X. 2019a. Multi-label Co-regularization for Semi-supervised Facial Action Unit Recognition. In *Advances in Neural Information Processing Systems*, 907–917.

Niu, X.; Han, H.; Yang, S.; Huang, Y.; and Shan, S. 2019b. Local Relationship Learning With Person-Specific Shape Regularization for Facial Action Unit Detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shao, Z.; Liu, Z.; Cai, J.; and Ma, L. 2018. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 705–720.

Shao, Z.; Liu, Z.; Cai, J.; Wu, Y.; and Ma, L. 2019. Facial action unit detection using attention and relation learning. *IEEE Transactions on Affective Computing* .

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12026–12035.

Simon, T.; Nguyen, M. H.; De La Torre, F.; and Cohn, J. F. 2010. Action unit detection with segment-based svms. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2737–2744. IEEE.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Song, T.; Liu, S.; Zheng, W.; Zong, Y.; and Cui, Z. 2020. Instance-Adaptive Graph for EEG Emotion Recognition. In *AAAI*, 2701–2708.

Song, T.; Zheng, W.; Song, P.; and Cui, Z. 2018. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing* .

Tang, C.; Zheng, W.; Yan, J.; Li, Q.; Li, Y.; Zhang, T.; and Cui, Z. 2017. View-independent facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 878–882. IEEE.

Tu, C.-H.; Yang, C.-Y.; and Hsu, J. Y.-j. 2019. IdenNet: Identity-Aware Facial Action Unit Detection. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1–8. IEEE.

Vaida, M.; and Patil, P. 2020. Semi-Supervised Graph Neural Network with Probabilistic Modeling to Mitigate Uncertainty. In *Proceedings of the 2020 the 4th International Conference on Information System and Data Mining*, ICISDM 2020, 152–156. New York, NY, USA: Association for Computing Machinery. ISBN 9781450377652. doi: 10.1145/3404663.3404680. URL https://doi.org/10.1145/3404663.3404680.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* .

Walecki, R.; (Oggi) Rudovic, O.; Pavlovic, V.; Schuller, B.; and Pantic, M. 2017. Deep Structured Learning for Facial Action Unit Intensity Estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, S.; Ding, H.; and Peng, G. 2020. Dual Learning for Facial Action Unit Detection Under Nonfull Annotation. *IEEE Transactions on Cybernetics* .

Wang, S.; Gan, Q.; and Ji, Q. 2017. Expression-assisted facial action unit recognition under incomplete au annotation. *Pattern Recognition* 61: 78–91.

Wang, S.; Hao, L.; and Ji, Q. 2018. Facial action unit recognition and intensity estimation enhanced through label dependencies. *IEEE Transactions on Image Processing* 28(3): 1428–1442.

Wang, Z.; Li, Y.; Wang, S.; and Ji, Q. 2013. Capturing global semantic relationships for facial action unit recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 3304–3311.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*.

Zhang, X.; Yin, L.; Cohn, J. F.; Canavan, S.; Reale, M.; Horowitz, A.; and Liu, P. 2013. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1–6. IEEE.

Zhang, Y.; Pal, S.; Coates, M.; and Ustebay, D. 2019. Bayesian graph convolutional neural networks for semi-supervised classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5829–5836.

Zhang, Y.; Zhao, R.; Dong, W.; Hu, B.-G.; and Ji, Q. 2018. Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7034–7043.

Zhao, K.; Chu, W.-S.; and Zhang, H. 2016. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3391–3399.