Knowledge Augmented Deep Neural Networks for Joint Facial Expression and Action Unit Recognition

Zijun Cui
Rensselaer Polytechnic Institute
cuiz3@rpi.edu

Tengfei Song
Southeast University
songtf@seu.edu.cn

Yuru Wang
Northeast Normal University
wangyr915@nenu.edu.cn

Qiang Ji
Rensselaer Polytechnic Institute
qji@ecse.rpi.edu

Abstract

Facial expression and action units (AUs) represent two levels of descriptions of the facial behavior. Due to the underlying facial anatomy and the need to form a meaningful coherent expression, they are strongly correlated. This paper proposes to systematically capture their dependencies and incorporate them into a deep learning framework for joint facial expression recognition and action unit detection. Specifically, we first propose a constraint optimization method to encode the generic knowledge on expression-AUs probabilistic dependencies into a Bayesian Network (BN). The BN is then integrated into a deep learning framework as a weak supervision for an AU detection model. A data-driven facial expression recognition(FER) model is then constructed from data. Finally, the FER model and AU detection model are trained jointly to refine their learning. Evaluations on benchmark datasets demonstrate the effectiveness of the proposed knowledge integration in improving the performance of both the FER model and the AU detection model. The proposed AU detection model is demonstrated to be able to achieve competitive performance without AU annotations. Furthermore, the proposed Bayesian Network capturing the generic knowledge is demonstrated to generalize well to different datasets.

1 Introduction

Facial expression is a key signal of human emotion. From the facial expression analysis perspective, there are two levels of expression descriptors: the global facial expression and the local Facial Action Units(AUs). These two descriptors lead to two research topics: Facial Expression Recognition(FER) and Facial Action Units Detection. AU is defined as facial muscle movements that correspond to a displayed expression according to Facial Action Coding System(FACS)[7]. AU activation is usually subtle and hard to annotate, thus the annotated AU data is limited and error prone. In comparison, the expression is global and easier to label. In addition, for both AU detection and FER problems, the data-driven models trained within datasets may generalize poorly to other datasets.

To learn a more generalizable model with limited AU annotation data, domain knowledge are considered. Some work used the manually designed knowledge directly from FACS [7] or muscle knowledge([23, 54]), whereas others constructed the dependencies among AUs from data ([48, 56, 52]) which are usually represented as a structural model (e.g. tree[14], graph[20], and graphical model[48, 45]). AU detection is then carried out by label propagation[23] or model training[44] using the captured knowledge. Though the relationships among AUs are included, the knowledge is based on local AU labels. Furthermore, the knowledge learned from a specific dataset cannot

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

generalize well. Since AUs and expressions are different levels of descriptors, they are closely related. In other words, they can complement each other and improve each other's performance through their interactions. Enlightened by this idea, expressions are employed as supplementary supervision for AU detection task([25, 26]). Besides expressions, facial landmarks are also considered as global information for AU detection[2]. On the other hand, AUs are also employed to enhance the FER tasks. Khorrami et al [16] showed that salient AU features can be obtained from deep neural networks that are trained for FER tasks. By incorporating AU information as domain knowledge into FER, performance improvements for FER can be achieved.

In this paper, we propose to perform joint AU detection and FER within a deep learning framework by leveraging the generic knowledge. Bayesian Network(BN) is employed to capture the generic knowledge on relationships among AUs and expression. Specifically, we propose to learn a BN purely from probability constraints derived from the generic knowledge and formulate the BN learning as a constraint optimization problem. The BN is then embedded into a deep learning framework to weakly supervise the learning of an AU detector. FER and AU detection modules are further jointly trained iteratively to improve each other's learning performance. By simultaneously leveraging both the AU-expression knowledge and the data, as well as integrating the knowledge via the interactions between AU detector and FER model, our FER model achieves better performance than a pure data-driven model, and our AU detector can generalize well to different datasets, and, more importantly, achieve comparable performance to existing supervised methods without any AU annotations.

2 Related Works

Facial expression recognition: Recent research on facial expression recognition leverages deep neural networks, such as AlexNet[19] and VGG[41], to obtain powerful representations. Pre-training is usually applied([33, 21, 4]). Pre-trained models can be constructed from non-facial data([4]) and other facial datasets[50]. Different approaches for improving FER performance are applied including data augmentation([1, 28]), network ensemble with diversified input[17] and network structures[38]. In addition, Tian et al [24] proposed to employ AU co-occurrence results for facial expression recognition. Probabilistic dependencies among expression-related factors are considered in [42, 11] for facial expression recognition. Attention mechanism has also been introduced to improve FER performance([32, 49]). Furthermore, different facial analysis tasks are closely related, and multi-task learning approach has been considered([15, 39]). Kollias et al [18] proposed a single network for joint AU detection and expression recognition. These methods are data-driven without considering domain knowledge and therefore do not generalize well to other datasets/subjects. Chen et al [4] proposed a facial motion prior network by considering the facial muscle movements corresponding to expressions. However, their facial muscle motion mask is computed within the dataset instead of from generic prior knowledge. In this work, we propose to encode the generic prior knowledge on the probabilistic dependencies among expressions and AUs into a Bayesian Network and integrate the prior knowledge into a deep neural network for improved FER.

Weakly supervised AU detection: AU annotation is challenging. Noise and errors exist in AU labels. Therefore, instead of solely relying on supervision from the AU annotations, many works seek supervisory information from other sources. Zhang et al [55] proposed context-aware attention mechanism for joint AU intensity estimation. Li et al [22] proposed to capture facial muscle movements for self-supervised AU detection. As the activation of AUs are produced by muscle contractions, strong dependencies exist among AUs. Hence, many works tried to encode AU dependencies as prior knowledge to improve AU detection under limited ground truth AU labels. Different methods are proposed to capture the relationships among AUs, including the kernel learning [52], LSTM [35] and graph convolutional network [34]. Moreover, graphical models attract much attention due to their power of modeling probabilistic dependencies([47, 43, 5]). In particular, Corneanu et al [5] appended the Conditional Random Field(CRF) to the end of a deep neural network to perform a message passing algorithm between AUs. There are other works leveraging not only the dependencies among AUs, but also their relationships to expressions([40, 45, 37, 48, 46]). In addition, Jyoti et al [13] applied expression features to enhance the AU detection performance. The knowledge mentioned above is extracted from a specific dataset, which limits their generalization ability to other datasets. Instead, Zhang et al[54] proposed to encode supervisory information from generic knowledge into loss terms for AU classifier training. Different from most of the existing works obtaining the knowledge from specific datasets, we propose to adapt the generic knowledge. Our model hence can generalize to different datasets/subjects. We propose to encode the generic knowledge with a BN which is different from [54]. Furthermore,we propose to embed the BN into a deep learning framework to perform joint AU detection and expression recognition.

3 Proposed Method

As shown in Figure 1, the proposed framework consists of a knowledge model in the middle represented by a Bayesian Network(BN) and three neural network based sub-models. The image-based FER model f_ψ on the top performs facial expression classification directly from image data. The AU model g_φ performs AU detection from the images, and the knowledge model K is used to weakly supervise the learning of the AU detector g_φ without requiring any AU annotations. The AU-based FER model h_ϕ in the right bottom performs expression recognition from AU detection results and is introduced mainly to assist the proposed model integration process. The three neural network models are initially trained independently and they are then refined jointly until convergence. In the end, we obtain the proposed FER model f_ψ and AU detection model g_φ. We first introduce each proposed model separately, and then the proposed model integration for joint training.

Figure 1: Overview of the proposed framework. Dotted lines represent back-propagation steps for each module.

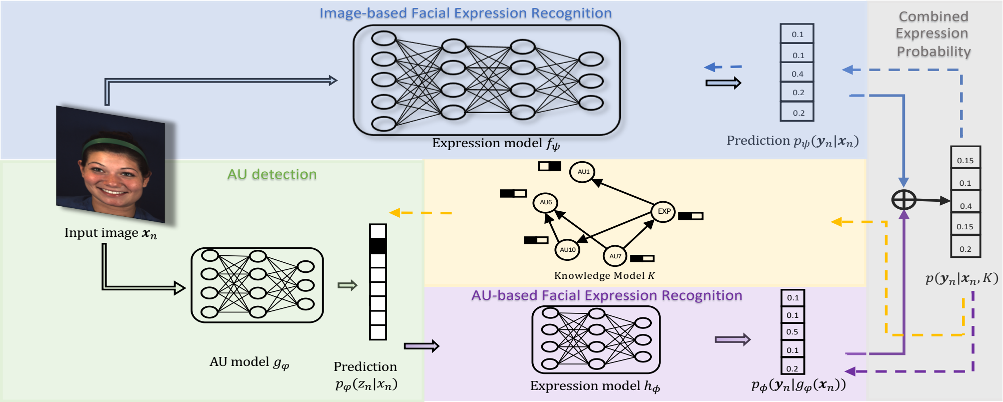
Image-based Facial Expression Recognition
Combined Expression Probability
0.1
0.1
0.4
0.2
0.2
Expression model f_ψ
Prediction $p_\psi(y_n|x_n)$
AU detection
Input image x_n
AU1
AU6
EXP
AU10
AU7
Knowledge Model K
0.15
0.1
0.4
0.15
0.2
$p(y_n|x_n, K)$
AU-based Facial Expression Recognition
0.1
0.1
0.5
0.1
0.2
AU model g_φ
Prediction $p_\varphi(z_n|x_n)$
Expression model h_ϕ
$p_\phi(y_n|g_\varphi(x_n))$

3.1 Knowledge-supervised AU Detection

We first discuss the generic knowledge extracted from the existing anatomic and psychological studies on facial expression generation mechanisms. We then show the knowledge encoding with a Bayesian Network(BN). Thirdly, we show the knowledge-supervised AU detection given the learned BN.

Generic knowledge as probabilities: We adapt the generic knowledge from existing studies that are applicable to different datasets. Generic knowledge is expressed as probabilities. Expression is denoted as $X^e = \{1, 2, ..., E\}$, where E is the total number of expression categories. AUs are denoted as $\{X^{au}_m\}_{m=1}^M$ where M is the total number of AUs and $X^{au}_m \in \{0, 1\}$. Inspired by [54], the generic knowledge is categorized into three types: expression-dependent single AU probabilities, expression-dependent joint AU probabilities, and expression-independent joint AU probabilities.

Table 1: Single AU in expressions from FACS[7] and [6](*in parentheses*).

AU	1	2	4	6	7	12	15	17
Anger			$P(\geq 0.7)$		$S(\geq 0.7)$			$S(0.52)$
Disgust			(0.31)					$S(\geq 0.7)$
Fear	$P(\geq 0.7)$	$P(0.57)$	$P(\geq 0.7)$		P			
Happy				$P(0.51)$	S	$P(\geq 0.7)$		
Sad	$P(0.6)$		$S\ (\geq 0.7)$	$S(0.5)$	P		$P(\geq 0.7)$	$S(0.67)$
Surprise	$P(\geq 0.7)$	$P(\geq 0.7)$						

1) For expression-dependent single AU probabilities, two sources are considered. According to FACS, given an expression, AUs can be grouped into *primary*(P) and *secondary*(S) categories as shown in Table 1. The primary AUs are the most expressive AUs with respective to the expression, and the secondary AUs may co-occur with primary AUs providing additional supports for the expression. Given a specific expression, the probability for its primary AU to be present is higher than its absence. For example, AU4(X^{au}_4) is a primary AU given the Anger

expression, and we have

$$p(X_4^{au}=1|X^e=\text{Anger}) > p(X_4^{au}=0|X^e=\text{Anger}) \tag{1}$$

AU that is neither primary nor secondary has higher chance for absence than occurrence. Besides, Du et al [6] quantitatively analyzed the relationships among expression and AUs based on their studies on different subjects and reported the probabilities for variant AUs under each expression. We include the reported probabilities under 6 basic expressions as another source of the generic knowledge as summarized in Table 1(Detailed probability formulations are in Appx.A);

Table 2: AU combinations from EMFACS[9]

Expression	AU
Anger	4+5, 4+7, 4+5+7, 17+24
Fear	1+2+4
Happy	6+12, 7+12
Sad	1+4, 6+15, 11+15, 11+17
Surprise	1+2+5, 1+2+26, 1+2+5+26

2) For expression-dependent joint AU probabilities, we consider two sources. According to FACS, given an expression, its primary AUs are more likely to be present than secondary AUs, and its secondary AUs have larger chance to appear than its other AUs. Secondly, the Emotional Facial Action Coding System(EMFACS) proposed by Wallace et al [9] studied the dependencies between combinations of AUs and expressions. We collect the AU combinations under basic expressions from EMFACS(Table 2)[1]. AUs within the same combination are likely to present together and are positively correlated. We formulate the probabilities by considering the pairwise positive correlation for each pair of AUs (X_i^{au}, X_j^{au}) within a AU combination(See Appx.A for details). For example, AU6 and AU12(X_6^{au}, X_{12}^{au}) are positively correlated given the Happy expression, i.e.,

$$\begin{aligned} p(X_6^{au}=1|X_{12}^{au}=1, X^e=\text{Happy}) > p(X_6^{au}=0|X_{12}^{au}=1, X^e=\text{Happy}) \\ p(X_6^{au}=1|X_{12}^{au}=1, X^e=\text{Happy}) > p(X_6^{au}=1|X_{12}^{au}=0, X^e=\text{Happy}) \end{aligned} \tag{2}$$

Table 3: AU correlations from anatomy

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

3) For expression-independent joint AU probabilities, we consider the dependencies among AUs caused by underlying facial muscle mechanism. For example, AU12(lip corner puller) and AU15(lip corner depressor) cannot show up together as their corresponding muscle groups(*Zgomaticus major* and *Depressor anguli oris* respectively) are unlikely to be activated simultaneously. The dependencies are further divided into positive correlations and negative correlations as summarized in Table 3. We formulate the pairwise dependencies for positively correlated AU pairs (X_i^{au}, X_j^{au}) as,

$$p(X_i^{au}=1|X_j^{au}=1) > p(X_i^{au}=0|X_j^{au}=1); \quad p(X_i^{au}=1|X_j^{au}=1) > p(X_i^{au}=1|X_j^{au}=0) \tag{3}$$

Similarly, we have the pairwise dependencies for negatively correlated AU pairs. The detailed formulations of probabilities based on the generic knowledge can be found in Appx.A.

BN learning with probability constraints: A BN is a direct acyclic graph (DAG) $\mathcal{G}=(\mathcal{V},\mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ denote nodes and edges respectively. BN can efficiently and accurately encode the joint distribution of AUs and expression. Instead of learning the BN from a specific dataset, we propose to learn a BN with probability constraints derived from the generic knowledge and formulate the BN learning as a constraint optimization problem. We employ the regression Bayesian Network(rBN). Consider a rBN with N nodes $\{X_i\}_{i=1}^N$ and the conditional probability of node X_i is defined as,

$$p(X_i=k|\pi(X_i)) = \sigma_M(\textstyle\sum_{j=1}^J w_{ijk}\pi_j(X_i) + b_{ik}) \tag{4}$$

where $X_i=\{k\}_{k=1}^K$ and $\pi_j(X_i)$ is the value of the j^{th} parent of the node X_i. w_{ijk} is the weight of j^{th} parent for node $X_i=k$, and b_{ik} is the bias for node $X_i=k$. $\sigma_M(x)$ is the softmax function. Specifically, $\{X_i\}_{i=1}^N = \{X^e, X_1^{au}, ..., X_M^{au}\}$. With weights $\boldsymbol{w}=\{w_{ijk}\}$, the structure of rBN is parameterized as a weighted adjacency matrix [2] $A \in R^{N\times N}$ with $A_{ij}=\sum_{k=1}^K ||w_{ijk}||_2^2$([58]). The structure is acyclic if and only if tr$(e^{A(\boldsymbol{w})\circ A(\boldsymbol{w})}) - N = 0$([57]). Learning a rBN is to learn $\boldsymbol{w}=\{w_{ijk}\}$ and $\boldsymbol{b}=\{b_{ik}\}$. Instead of learning a rBN with training data, we consider learning a rBN with probability constraints. We categorize probability constraints into three groups for calculation: strictly inequality constraints $\{g_i(\boldsymbol{w},\boldsymbol{b}) < 0\}_{i=1}^G$, inequality constraints $\{l_j(\boldsymbol{w},\boldsymbol{b}) \leq 0\}_{j=1}^L$ and

[1] Disgust expression doesn't have the corresponding most likely AU combinations and thus is not included.

[2] $A_{ij}=0$ indicates that there is no link pointing from node X_j to node X_i

equality constraints $\{h_k(\boldsymbol{w}, \boldsymbol{b}) = 0\}_{k=1}^{H}$. We follow Eq.1-3 to derive probability constraints given the generic knowledge. Most of the constraints belong to strictly inequality constraints(See Appx.B for details). Additional variables s_j are introduced to handle strictly inequality constraints. In particular, we apply the exponential function to define positive margins given variables s_i, and each of the strictly inequality constraints becomes $g_i(\boldsymbol{w}, \boldsymbol{b}) + e^{s_i} = 0$. Each constraint imposes a non-linear constraint on the joint probability distribution of the expression and AUs. We employ the penalty method [51] and treat each probability constraint derived from the generic knowledge as a soft constraint. A penalty function $f(\boldsymbol{w}, \boldsymbol{b}; \boldsymbol{s})$ is then defined measuring the violation of constraints given current margins defined by $e^{\boldsymbol{s}}$, i.e.,

$$f(\boldsymbol{w}, \boldsymbol{b}; \boldsymbol{s}) = \tfrac{1}{G} \textstyle\sum_{i=1}^{G} \log((g_i(\boldsymbol{w}, \boldsymbol{b}) + e^{s_i})^2 + 1) + \tfrac{1}{L} \textstyle\sum_{j=1}^{L} \log((l_j^+(\boldsymbol{w}, \boldsymbol{b}))^2 + 1) + \tfrac{1}{K} \textstyle\sum_{k=1}^{K} \log((h_i(\boldsymbol{w}, \boldsymbol{b}))^2 + 1) \quad (5)$$

where $l_j^+(\boldsymbol{w}, \boldsymbol{b}) = \max\{0, l_j(\boldsymbol{w}, \boldsymbol{b})\}$. $f(\boldsymbol{w}, \boldsymbol{b}; \boldsymbol{s}) = 0$ if and only if all the constraints are satisfied given current margin $e^{\boldsymbol{s}}$, and $f(\boldsymbol{w}, \boldsymbol{b}; \boldsymbol{s}) > 0$ otherwise. Learning a rBN with probability constraints is then formulated as a constraint optimization problem,

$$\begin{gathered} \boldsymbol{w}^*, \boldsymbol{b}^*, \boldsymbol{s}^* = \arg \min_{\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{s}} f(\boldsymbol{w}, \boldsymbol{b}; \boldsymbol{s}) + \gamma ||\boldsymbol{w}||_1 - \mu ||\boldsymbol{s}||_2^2 \\ \text{s.t. tr}(e^{A(\boldsymbol{w}) \circ A(\boldsymbol{w})}) = 0 \end{gathered} \quad (6)$$

where $||\boldsymbol{w}||_1$ is the L1-norm of $\boldsymbol{w}$ penalizing the density of the structure and $||\boldsymbol{s}||_2^2$ is the squared L2-norm of $\boldsymbol{s}$ encouraging the bigger positive margins. μ is set to be $\frac{1}{G}$. $\boldsymbol{w}, \boldsymbol{b}$ and $\boldsymbol{s}$ are updated iteratively until convergence. The learned rBN with $\boldsymbol{w}^*$ and $\boldsymbol{b}^*$ is denoted as the knowledge model K. The visualization of the structure and empirical analysis on the learned BN are in Appx.C. Our work differs from [54] in terms of the approach for knowledge encoding. In [54], each constraint is encoded as a loss term independently. In contrast, we apply the BN so that not only individual constraints but also the underlying structured dependencies among AUs and expression are captured comprehensively and systematically.

AU detection model: The AU detection model g_φ(in the middle of the Fig. 1) takes a facial image $\boldsymbol{x}_n$, $n = 1, 2, ..., N$ as input and outputs the probability for AUs, i.e. $p(\boldsymbol{z}_n | \boldsymbol{x}_n)$, where $\boldsymbol{z}_n$ represents the AU configuration. For each configuration $\boldsymbol{z}_n$, we compute a cross-entropy loss $l(\boldsymbol{z}_n, g_\phi(\boldsymbol{x}_n))$, and we consider the expected cross-entropy over $p(\boldsymbol{z}_n | \boldsymbol{y}_n^{GT}, K)$ to obtain the model parameters φ,

$$\varphi^* = \arg\min_\varphi \textstyle\sum_{n=1}^{N} E_{p(\boldsymbol{z}_n | \boldsymbol{y}_n^{GT}, K)} l(\boldsymbol{z}_n, g_\varphi(\boldsymbol{x}_n)) \quad (7)$$

where $p(\boldsymbol{z}_n | \boldsymbol{y}_n^{GT}, K)$ is computed from the BN model given the facial expression label $\boldsymbol{y}_n^{GT}$.

3.2 Facial Expression Recognition Models

Image-based FER model: The image-based FER model f_ψ (the top component in Fig.1) takes a facial image as input and outputs the probabilities for facial expression classes, $p_\psi(\boldsymbol{y}_n | \boldsymbol{x}_n)$. The training consists of the input images $\boldsymbol{x}_n$, $n = 1, 2, ..., N$, and their expression labels $\boldsymbol{y}_n^{GT}$. Cross-entropy loss (denoted as l) is used to train the deep model to obtain the model parameters ψ :

$$\psi^* = \arg\min_\psi \tfrac{1}{N} \textstyle\sum_{n=1}^{N} l(\boldsymbol{y}_n^{GT}, f_\psi(\boldsymbol{x}_{\boldsymbol{n}})) \quad (8)$$

AU-based FER model: The AU-based FER model performs expression recognition using the AU detection results and can indirectly capture the AU-expression relationships in the knowledge model. Such a model is practically useful in assisting the model integration as it produces better FER performance than directly using the knowledge model K(Detailed analysis are in Appx.D). Specifically, the AU-based FER model h_ϕ takes the output of the AU detector $g_\varphi(\boldsymbol{x}_n)$ as its input and generates the probability for each facial expression class $p_\phi(\boldsymbol{y}_n | g_\varphi(\boldsymbol{x}_n), K)$. Standard cross-entropy loss function(denoted as l) is used to obtain its parameters ϕ as follows:

$$\phi^* = \arg\min_\phi \tfrac{1}{N} \textstyle\sum_{n=1}^{N} l(\boldsymbol{y}_n^{GT}, h_\phi(g_\varphi(\boldsymbol{x}_n))) \quad (9)$$

The combined FER model: The image-based and AU-based FER models produce independent facial expression recognition results, i.e., $p_\psi(\boldsymbol{y}_n | \boldsymbol{x}_n)$ and $p_\phi(\boldsymbol{y}_n | g_\varphi(\boldsymbol{x}_n), K)$ respectively. Their results are then combined to produce the expression probability $p(\boldsymbol{y}_n | \boldsymbol{x}_n, K)$, i.e.,

$$p(\boldsymbol{y}_n | \boldsymbol{x}_n, K) = w_1 p_\psi(\boldsymbol{y}_n | \boldsymbol{x}_n) + w_2 p_\phi(\boldsymbol{y}_n | g_\varphi(\boldsymbol{x}_n), K) \quad (10)$$

where w_1 and w_2 are the weights. Entropy is applied to quantify the weights, i.e., $w_1 = \sigma_M(-h_1)$ and $w_2 = \sigma_M(-h_2)$ where $h_1 = H_{p_\psi(\boldsymbol{y}_n | \boldsymbol{x}_n)}(\boldsymbol{y}_n | \boldsymbol{x}_n)$ and $h_2 = H_{p_\phi(\boldsymbol{y}_n | g_\varphi(\boldsymbol{x}_n), K)}(\boldsymbol{y}_n | \boldsymbol{x}_n)$. $\sigma_M(x_i) =$

$\frac{e^{x i}}{\sum_{i=1}^{2} e^{x i}}$ is the softmax function. The lower the entropy is, the higher the weight is. The combined expression distribution is then consistent with both data information and the underlying knowledge.

3.3 AU and Expression Models Integration

We have introduced the AU detection model and the FER model thus far. They are learnt independently. Since the expression and AUs are highly correlated, it makes sense to perform their joint recognition by exploiting their dependencies. To this end, we propose to augment the loss functions for AU detection model(Eq. 7) and image-based FER model(Eq. 8) via the combined expression distribution(Eq. 10) to perform the joint expression recognition and AU detection.

Expression-augmented AU detection model: We incorporate the combined expression probability $p(\boldsymbol{y}_n|\boldsymbol{x}_n, K)$ into AU detection model. In particular, we introduce a regularization term to the AU loss function (Eq. 7) by considering the expected loss over $p(\boldsymbol{y}_n|\boldsymbol{x}_n, K)$ as

$$\varphi^* = \arg\min_\varphi \textstyle\sum_{n=1}^N E_{p(\boldsymbol{z}_n|\boldsymbol{y}_n^{GT}, K)} l(\boldsymbol{z}_n, g_\varphi(\boldsymbol{x}_n)) + \lambda_1 E_{p(\boldsymbol{y}_n|\boldsymbol{x}_n, K)} E_{p(\boldsymbol{z}_n|\boldsymbol{y}_n, K)} l(\boldsymbol{z}_n, g_\varphi(\boldsymbol{x}_n)) \quad (11)$$

where λ_1 is a hyper-parameter to be tuned. Through the regularization term, the expression recognition results are integrated into the AU detection model.

Knowledge-augmented image-based FER model: The interactions between AUs and facial expression are both way. Given the combined distribution $p(\boldsymbol{y}_n|\boldsymbol{x}_n, K)$, we introduce a regularization term by considering the expected loss over $p(\boldsymbol{y}_n|\boldsymbol{x}_n, K)$ to the loss function(Eq. 8) to augment FER,

$$\psi^* = \arg\min_\psi \tfrac{1}{N} \textstyle\sum_{n=1}^N l(\boldsymbol{y}_n^{GT}, f_\psi(\boldsymbol{x_n})) + \lambda_2 E_{p(\boldsymbol{y}_n|\boldsymbol{x}_n, K)} l(\boldsymbol{y}_n, f_\psi(\boldsymbol{x_n})) \quad (12)$$

λ_2 is a hyper-parameter to be tuned. Through the regularization term, the AU detection results and the AU-expression relationships encoded in the knowledge model are integrated into the FER model.

Through Eq. 11 and Eq. 12, we can systematically combine the AU detection model and the FER model. These models interact with each other during training to improve each other's performance. Furthermore, we apply an iterative updating procedure that can continuously update each model until convergence: given an updated combined distribution, update the image-based FER model and the AU model. The AU-based FER model is updated accordingly given the updated AU model. The pseudo-code for the proposed model integration for joint training is summarized in Algorithm 1.

Algorithm 1 Iterative model training

```
Input:
    Training Data $\mathcal{D} = \{\boldsymbol{x}_n, \boldsymbol{y}_n^{GT}\}_{n=1}^N$
    Knowledge model $K$
    Hyper-parameters $\lambda_1$, $\lambda_2$
Output:
    Image-based expression model $f_\psi$
    AU detection model $g_\varphi$
Initialize $f_\psi$ via pre-train with Eq. 8
Initialize $g_\varphi$ via pre-train with Eq. 7
Initialize $h_\phi$ via pre-train with Eq. 9
while Not Converging do
    Apply $f_\psi$ to predict $p_\psi(\boldsymbol{y}_n|\boldsymbol{x}_n)$
    Apply $g_\varphi$ to predict $p_\varphi(\boldsymbol{z}_n|\boldsymbol{x}_n)$
    Apply $h_\phi$ to predict $p_\phi(\boldsymbol{y}_n|g_\varphi(\boldsymbol{x}_n), K)$
    Combine probability with Eq. 10
    Update $f_\psi$ with Eq. 12
    Update $g_\varphi$ with Eq. 11
    Update $h_\phi$ with Eq. 9
end while
```

4 Experiments

Databases: We consider four benchmark datasets: BP4D-Spontaneous database[53], Extended CohnKanande(CK+) database[29], M&M Initiative(MMI) database[36] and EmotioNet[31]. The BP4D [53] is a spontaneous database containing 328 sequences from 41 subjects. Each sequence is labelled with one expression category and many frames along the sequence contain neutral status without expression. As our method requires the presence of expression, we collect 803 apex frames in total. The CK+ [29] is a posed expression database that contains 309 sequences from 109 subjects. For each sequence, expression starts from neutral intensity to the strongest intensity for a specific expression category. Typically, the last frame of each sequence is extracted. In total, 309 frames are collected. MMI [36] is a posed expression database. 238 clips of 28 subjects are collected from Part II of MMI. Typically, three frames around the center of each sequence are selected. In total, 504 frames are collected from labelled sequences with frontal face. For BP4D, CK+, and MMI, annotations for both AUs and 6 basic expressions are collected. The EmotioNet[31] is collected in the wild. Annotations in EmotioNet are noisy as they are automatically generated by existing algorithms. All 24,556 images are collected and around 2,000 images have annotations for both AUs and expressions, including compound expressions. Due to occlusion, AU annotations are incomplete, and thus we only consider expressions. In total, 537 images with 6 basic expressions are collected(Appx.E for statistical information of datasets). We employ 5-fold subject-independent cross-validation experiments.

Models and hyperparameters: Three-layer CNN is applied for the AU detection model. The kernel size of each layer is 5x5, 4x4, 3x3 respectively. Three fully connected layers are adopted for the AU-based FER model. VGG-19[41] is applied for image-based FER model and is pre-trained with FER2013[10](Appx.F for analysis on the effects of pre-training). AdamOptimizer is applied with learning rate 0.0005. $\gamma = 0.001$(Eq.9). λ_1 = 0.005(Eq.11), and λ_2=0.001 (Eq.12). Values of λ_1 and λ_2 are selected based on grid search from the range $\{0.0005, 0.001, 0.005, 0.01, 0.5, 1\}$(Appx.G for analysis on effects of λ_1 and λ_2).

4.1 Action Units Detection

We evaluate the performance of the proposed AU detection model. We denote the initial AU detection model as AUD-BN(Eq.7) learned with the generic BN and the expression-augmented AU detection model as AUD-EA(Eq.11). We firstly evaluate the AUD-BN to demonstrate the effectiveness of the generic knowledge. For comparison, we train the AU detection model with GT AU annotations provided by the dataset(denoted as AUD-GT). We then compare the performance of the AUD-BN and AUD-EA to demonstrate the effectiveness of the model integration. Finally, we compare the proposed AU detection model to the state-of-the-art models. F1-score is applied as the evaluation metric. We report the F1-score for each AU and the averaged F1-score over AUs.

4.1.1 AU Detection Evaluation

Effectiveness of the generic BN: To evaluate the effectiveness of the proposed generic BN(gBN), we consider the performance of AU detection with the gBN and a subset of GT AU annotations. For the training samples without GT AU annotations, we apply Eq.7. For the rest with GT AU annotations, an additional regularization term is introduced to Eq.7 calculating the cross-entropy loss between GT AU annotations and predicted AU labels. The coefficient of the regularization term is set to be 1. We consider BP4D and CK+ for evaluation. Results are shown in Table 4. For CK+, gBN with 80% GT AU annotations achieves average F1-score 0.79, better than AUD-GT(0.78). Furthermore, gBN with 100% AU annotations achieves 0.83, the best performance. For BP4D, with gBN and only 60% GT AU annotations, it achieves the same performance as the performance using all GT annotations. These results demonstrate that gBN can capture additional knowledge beyond the labels and they can apply to different datasets. And by leveraging the gBN, the AU detection model has less dependency on GT annotations, and thus is more data efficient.

Table 4: Evaluation of generic BN(gBN) with a subset of GT AU annotations

Database	BP4D									CK+								
AU	1	2	4	6	7	12	15	17	Avg	1	2	4	6	7	12	15	17	Avg
gBN + 0%GT	.54	.44	.54	.60	.80	.55	.42	.62	.56	.79	.78	.72	.69	.44	.90	.47	.75	.69
gBN + 20%GT	.54	.36	.53	.64	.81	.65	.38	.62	.57	.83	.88	.71	.64	.49	.85	.53	.74	.71
gBN + 40%GT	.46	.35	.52	.74	.81	.76	.42	.65	.59	.85	.84	.75	.77	.52	.90	.49	.75	.73
gBN + 60%GT	.48	.38	.51	.78	.82	.79	.50	.71	.62	.90	.88	.78	.75	.59	.92	.51	.84	.77
gBN + 80%GT	.64	.36	.61	.79	.84	.84	.38	.70	.64	.89	.93	.79	.78	.56	.92	.62	.79	.79
gBN + 100%GT	.54	.44	.58	.77	.82	.84	.51	.74	**.65**	.93	.93	.81	.78	.63	.92	.74	.86	**.83**
AUD-GT	.49	.36	.56	.78	.85	.83	.36	.70	.62	.87	.90	.85	.74	.53	.91	.60	.84	.78

Effectiveness of model integration: The results of AU detection with AUD-BN and AUD-EA are summarized in Table 5. For all three datasets, the AUD-EA achieves better performance compared to AUD-BN with the generic BN. In particular, the F1-score that AUD-EA achieves averaged over 8 AUs is 5% higher than AUD-BN for CK+ and 11% higher for MMI. In addition, the proposed AUD-EA with the generic BN achieves comparable performance with AUD-GT for BP4D and CK+. For MMI, we obtain the F1-score 54% with AUD-GT, and with AUD-EA, we achieve 58%, better than AUD-GT. The reason is that AU annotations in MMI are very unbalanced, in particular for AU6 and AU15 and thus AUD-GT fails to produce good performance. On the other hand, applying the generic knowledge through the model integration without AU annotations, our proposed AUD-EA produces the best performance. These results demonstrate that the proposed model integration is effective in improving the AU detection performance. In addition, the generic BN is further demonstrated to represent well the underlying generic knowledge on the expression-AUs relationships. The generic BN can supervise the training of the AU detector effectively and generalizes well to different datasets.

Table 5: Evaluation of AU detection models: AUD-BN and AUD-EA

	AUD-BN									AUD-EA									AUD-GT								
AU	1	2	4	6	7	12	15	17	Avg	1	2	4	6	7	12	15	17	Avg	1	2	4	6	7	12	15	17	Avg
BP4D	.54	.44	.54	.60	.80	.55	.42	.62	.56	.53	.36	.55	.68	.83	.57	.39	.65	.57	.49	.36	.56	.78	.85	.83	.36	.70	**.62**
CK+	.79	.78	.72	.69	.44	.90	.47	.75	.69	.90	.82	.67	.70	.46	.94	.68	.75	.74	.87	.90	.85	.74	.53	.91	.60	.84	**.78**
MMI	.59	.50	.59	.39	.40	.58	.27	.42	.47	.67	.66	.70	.40	.43	.87	.45	.44	**.58**	.58	.78	.74	.07	.52	.83	.31	.52	.54

4.1.2 Comparison to the state-of-the-art Methods

We compare the proposed expression-augmented AUD model(AUD-EA) with generic BN(gBN) to the state-of-the-art AU detection methods. To be clear, we focus our comparison on only AU detection and exclude the work on AU intensity estimation. We consider the state-of-the-art methods HTL [40], LP-SM [54] and TCAE [22] which apply AU-expression relationships as weak supervisions and don't require AU annotations. In addition, we compare to SoA supervised learning methods as JPML [56], HRBM [47], MC-LVM [8] and AU R-CNN [30]. We report the averaged F1-score over 8 AUs and the results are shown in Table 6. $(^*)$ indicates the reported results. As shown in Table 6, our proposed method outperforms all the other weakly supervised methods. For example, our proposed AUD-EA with generic BN produces 74% for CK+, which is 2% better than LP-SM and 7% better than HTL. In particular, TCAE is a sequence-based method leveraging temporal information for weak supervision, and our proposed approach achieves better performance compared to TCAE. Compared to the supervised learning methods, because our method doesn't require any annotation, our performance is not that competitive. For MMI, HRBM fails to perform well due to the unbalanced labels. On the other hand, by leveraging the generic knowledge and model integration, our AUD-EA achieves the best performance.

Table 6: Comparison to the SoAs on AU detection.

Supervision	Method	BP4D	CK+	MMI
Supervised	HRBM[47]	.67	.79	.56
	MC-LVM[8]	-	**.80***	-
	JPML[56]	**.68***	.78*	-
	AU R-CNN[30]	.63*	-	-
Weakly-supervised	HTL[40]	.50	.66	.42
	LP-SM[54]	.55	.72*	.50
	TCAE[22]	.56*	-	-
	AUD-BN(baseline)	.56	.69	.47
	AUD-EA(gBN)	**.57**	**.74**	**.58**

4.2 Facial Expression Recognition

We evaluate the performance of the proposed FER models. We denote the initial image-based FER model trained with GT expression annotations(Eq. 8) as FER-I, and the knowledge-augmented image-based FER model(Eq. 12) as FER-IK. We firstly evaluate the effectiveness of the proposed joint training by comparing the performance of FER-I and FER-IK. We then compare the proposed FER-IK to the state-of-the-art FER models. We apply classification accuracy as the evaluation metric.

Image-based FER evaluation: We compare the performance of FER-I and FER-IK to demonstrate the effectiveness of the model integration. Besides BP4D, CK+ and MMI, we also consider the EmotioNet(EmNet) to further demonstrate the effectiveness of the proposed model integration with the generic BN on noisy and challenging dataset. We report the expression accuracy for both FER-I and FER-IK with the generic BN as shown in Table 7. It is clear from the table that for all datasets, FER-IK significantly outperforms FER-I. In particular, the FER-IK achieves 22.14% accuracy improvements for BP4D. These results show that by integrating the generic knowledge through the joint training, the FER performance can be significantly improved. The proposed generic BN can apply to different datasets. In addition, for EmtioNet, FER-IK achieves significant improvement which demonstrate the effectiveness of our proposed approach on noisy and challenging datasets, such as EmotioNet.

Table 7: Evaluation of the FER model

Model	BP4D	CK+	MMI	EmNet
FER-I	61.68	94.29	67.35	80.85
FER-IK	**83.82**	**97.59**	**84.90**	**95.55**

Comparison to the state-of-the-art methods: We compare the performance of proposed knowledge-augmented image-based FER model(FER-IK) with generic BN(gBN) to the state-of-the-art methods. Results are shown in Table 8. (*) indicates reported results. For FMPN-FER[4][3] and DeepEmotion[32][4], we perform experiments with default hyperparameters suggested in the papers. Our proposed method outperforms all the other SoA FER methods. In particular, for BP4D and EmotioNet, we achieve significant improvement. BP4D is a spontaneous dataset, and we achieve 4.28% accuracy improvement compared to the DeepEmotion. EmotioNet is more challenging as its annotations are very noisy. And both the FMPN-FER and the DeepEmotion don't produce competitive performance for the EmotioNet. On the other hand, by incorporating the generic knowledge through model integration to compensate label errors, our model produces the outstanding performance. In addition, for FMPN-FER, we follow the procedure suggested by the paper and apply the prior facial motion mask obtained from CK+ to BP4D and EmotioNet. Its poor performance for BP4D and EmotioNet indicates that the prior mask obtained from a specific dataset can not generalize well to other datasets.

Table 8: Comparison with SoA FER methods

Methods	BP4D	CK+	MMI	EmotioNet
STM-Explet[27]	-	94.19*	75.12*	-
DTAGN(Joint)[12]	-	97.25*	70.24*	-
DeRL[50]	-	97.30*	73.23*	-
ILCNN[3]	-	94.35*	70.67*	-
DAM-CNN[49]	-	95.88*	-	-
FMPN-FER[4]	60.16	96.53	82.74*	84.88
DeepEmotion[32]	79.54	95.23	72.66	81.51
FER-I(baseline)	61.68	94.29	67.35	80.85
FER-IK(gBN)	**83.82**	**97.59**	**84.90**	**95.55**

5 Conclusion

This paper proposes a knowledge augmented deep learning framework for joint AU detection and facial expression recognition. We first propose a constraint optimization method to encode the generic knowledge on expression-AUs dependencies into a Bayesian Network (BN). We then embed the BN model into a deep learning framework to perform weakly supervised AU detection. We further introduce a joint training procedure to exploit the interactions between AU detection and FER for improved performance on both tasks. Experiments on benchmark datasets show that the proposed method achieves improved performance for both FER and AU detection. Specifically, for facial expression, the proposed knowledge-augmented FER model outperforms the SoA FER models. For AUs, the proposed AU detection model, trained with a generic BN without any GT AU annotations, significantly outperforms the SoA weakly-supervised methods and achieves comparable results to SoA supervised methods. Experiments also show that with the incorporation of the generic knowledge, our model generalizes well to different datasets, even perform well on the MMI dataset with unbalanced AU annotations and the challenging EmotioNet with noisy expression labels.

Acknowledgement

The work described in this paper is supported in part by a DARPA grant FA8750-17-2-0132, and in part by the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons).

Broader Impact

This work is focused on two computer vision tasks: facial expression recognition and facial action units detection. The potential broader impacts of this work are listed as follows:

Benefits: Facial expression recognition can benefit many applications, including HCI, social robotics, medical diagnosis, games animation, etc. By leveraging the domain knowledge, our proposed models have less dependence on training data and thus the data efficiency is improved. In other words, it may release domain experts from the heavy workload on labeling data. Furthermore, as the domain

[3] https://github.com/donydchen/FMPN-FER

[4] https://github.com/omarsayed7/Deep-Emotion

knowledge is generic, our propose model can generalize well to different datasets. Hence, given new datasets, additional training process is promised to be no longer necessary.

Risks: Facial expression recognition has some privacy concerns. For example, through facial expression recognition systems, individuals' emotional reactions to certain messages, news or figures can be tracked. Also individuals' emotional reactions to events can be monitored in the public places with facial expression recognition systems.

References

[1] Iman Abbasnejad, Sridha Sridharan, Dung Nguyen, Simon Denman, Clinton Fookes, and Simon Lucey. Using synthetic data to improve facial expression analysis with 3d convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[2] C. F. Benitez-Quiroz, Y. Wang, and A. M. Martinez. Recognition of action units in the wild with deep nets and a new global-local loss. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3990–3999, 2017.

[3] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James OReilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, May 2018.

[4] Yuedong Chen, Jianfeng Wang, Shikai Chen, Zhongchao Shi, and Jianfei Cai. Facial motion prior networks for facial expression recognition. *arXiv preprint arXiv:1902.08788*, 2019.

[5] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. Deep structure inference network for facial action unit recognition. In *Proceedings of European Conference on Computer Vision*, 2019.

[6] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[7] Paul Ekman, Wallace V. Friesen, and J. C. Hager. Facial action coding system. *A Human Face, Salt Lake City, UT*, 2002.

[8] 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.

[9] Wallace V. Friesen and Paul Ekman. Emfacs-7: Emotional facial action coding system. 1983.

[10] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

[11] Jun He, Xiaocui Yu, Lejun Yu, and Bo Sun. Facial emotion and action unit recognition based on bayesian network. In *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, pages 363–368, 2019.

[12] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991, 2015.

[13] Shreyank Jyoti, Garima Sharma, and Abhinav Dhall. Expression empowered residen network for facial action unit detection. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[14] Sebastian Kaltwang, Sinisa Todorovic, and Maja Pantic. Latent trees for estimating intensity of facial action units. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[15] Takuhiro Kaneko, Kaoru Hiramatsu, and Kunio Kashino. Adaptive visual feedback generation for facial expression improvement with multi-task deep neural networks. *ACM on Multimedia Conference*, 2016.

[16] Pooya Khorrami, Tom Le Paine, and Thomas S. Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[17] Bo-Kyeong Kim, Suh-Yeon Dong, Jihyeon Roh, Geonmin Kim, and Soo-Young Lee. Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, 2016.

[18] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior\a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019.

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

[20] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of AAAI Conference on Artificial Intelligence*,

2019.

[21] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

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

[23] Yongqiang Li, Jixu Chen, Yongping Zhao, and Qiang Ji. Data-free prior model for facial action unit recognition. *IEEE Transactions on Affective Computing*, 2013.

[24] Ying li Tian, Takeo Kanade, and Jeffrey F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001.

[25] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-aware deep networks for facial expression recognition. In *Proceedings of IEEE International Conference on Automatic Face*, 2013.

[26] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-inspired deep networks for facial expression feature learning. *Neurocomputing*, 2015.

[27] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756, 2014.

[28] Xiaofeng Liu, B.V.K Vijaya Kumar, Jane You, and Ping Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, 2017.

[29] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010.

[30] Chen Ma, Li Chen, and Junhai Yong. Au r-cnn: Encoding expert prior knowledge into r-cnn for action unit detection. *Neurocomputing*, 355:35–47, 2019.

[31] Iain Matthews and Simon Baker. Active appearance models revisited. *International journal of computer vision*, 60(2):135–164, 2004.

[32] Shervin Minaee and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *arXiv preprint arXiv:1902.01019*, 2019.

[33] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of ACM on International Conference on Multimodal Interaction*, 2015.

[34] 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 907–917, 2019.

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

[36] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *ICME*, 2005.

[37] Guozhu Peng and Shangfei Wang. Weakly supervised facial action unit recognition through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, 2018.

[38] Gerard Pons and David Masi. Supervised committee of convolutional neural networks in automated facial expression analysis. *IEEE Transactions on Affective Computing*, 2017.

[39] Gerard Pons and David Masip. Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. *arXiv preprint arXiv:1802.06664*, 2018.

[40] Adria Ruiz, Joost Van de Weijer, and Xavier Binefa. From emotions to action units with hidden and semi-hidden-task learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

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

[42] Maninderjit Singh, Anima Majumder, and Laxmidhar Behera. Facial expressions recognition system using bayesian inference. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 1502–1509. IEEE, 2014.

[43] Yale Song, Daniel McDuff, Deepak Vasisht, and Ashish Kapoor. Exploiting sparsity and co-occurrence structure for action unit recognition. In *Proceedings of IEEE International Conference on Automatic Face*, 2015.

[44] Robert Walecki, Ognjen (Oggi) Rudovic, Vladimir Pavlovic, Bjoern Schuller, and Maja Pantic. Deep structured learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition, 2017.

[45] Jun Wang, Shangfei Wang, and Qiang Ji. Facial action unit classification with hidden knowledge under incomplete annotation. In *ACM International Conference on Multimedia Retrieval*, 2015.

[46] Shangfei Wang, Guozhu Peng, and Qiang Ji. Exploring domain knowledge for facial expression-assisted action unit activation recognition. *IEEE Transactions on Affective Computing*, 2018.

[47] 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.

[48] Shan Wu, Shangfei Wang, Bowen Pan, and Qiang Ji. Deep facial action unit recognition and intensity estimation from partially labelled data. *IEEE Transactions on Attective Computing*, 2019.

[49] Siyue Xie, Haifeng Hu, and Yongbo Wu. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. *Pattern Recognition*, 92:177 – 191, 2019.

[50] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2168–2177, 2018.

[51] Özgür Yeniay. Penalty function methods for constrained optimization with genetic algorithms. *Mathematical and computational Applications*, 10(1):45–56, 2005.

[52] Xiao Zhang and Mohammad H. Mahoor. Task-dependent multi-task multiple kernel learning for facial action unit detection. *Pattern Recognition*, 2016.

[53] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *FG workshop*, 2013.

[54] 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.

[55] 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.

[56] Kaili Zhao, WenSheng 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.

[57] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pages 9472–9483, 2018.

[58] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425, 2020.