A 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. With 6 basic expressions, we have $E = 6$. AUs are denoted as $\{X_m^{au}\}_{m=1}^M$ where M is the total number of AUs and $X_m^{au} \in \{0, 1\}$, where 1 for presence and 0 for absence. The generic knowledge is categorized into three types: expression-dependent single AU probabilities, expression-dependent joint AU probabilities, and expression-independent joint AU probabilities.

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. 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 it's absence. For example, AU4(X_4^{au}) 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. For example, AU1(X_1^{au}) is neither a primary AU nor a secondary AU given the Anger expression, and we have

$$p(X_1^{au} = 1 | X^e = \text{Anger}) < p(X_1^{au} = 0 | X^e = \text{Anger}) \tag{2}$$

Besides, Du et al [5] studied both basic expressions and compound expressions. They quantitatively analyzed the relationships among expressions 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. For example, the probability for AU4(X_4^{au}) being present is 31% given a Disgust expression, and we have

$$p(X_4^{au} = 1 | X^e = \text{Disgust}) = 0.31 \tag{3}$$

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. For example, given a Sad expression, we have AU1(X_1^{au}) being a primary AU, AU4(X_4^{au}) being a secondary AU, and AU7(X_7^{au}) being a AU that is neither primary nor secondary, and we have

$$\begin{aligned} p(X_1^{au} = 1 | X^e = \text{Sad}) > p(X_4^{au} = 1 | X^e = \text{Sad}) \\ p(X_4^{au} = 1 | X^e = \text{Sad}) > p(X_7^{au} = 1 | X^e = \text{Sad}) \end{aligned} \tag{4}$$

Secondly, the Emotional Facial Action Coding System(EMFACS) studied the dependencies between combinations of AUs and expressions. We collect the AU combinations under basic expressions from EMFACS. 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. For example, AU6 and AU12(X_6^{au}, X_{12}^{au}) are positively correlated given the Happy expression, i.e.,

$$\begin{aligned} p(X_6^{au} = 1 | X_{12}^{au} = 1, X^e = \text{Happy}) > p(X_6^{au} = 0 | X_{12}^{au} = 1, X^e = \text{Happy}) \\ p(X_6^{au} = 1 | X_{12}^{au} = 1, X^e = \text{Happy}) > p(X_6^{au} = 1 | X_{12}^{au} = 0, X^e = \text{Happy}) \end{aligned} \tag{5}$$

The first equation in Eq.5 indicates that when AU12 appears, AU6 is more likely to appear than not under the Happy expression. The second indicates that the probability of the occurrence of

AU6 given the presence of AU12 is higher than when AU12 does not appear under the Happy expression;

3) For expression-independent joint AU probabilities, we consider the dependencies among AUs caused by underlying facial muscle mechanism. The dependencies are further divided into positive correlations and negative correlations. AUs that are likely to co-occur share the positive correlation. We formulate the pairwise dependencies for positively correlated AU pairs (X_i^{au}, X_j^{au}) as,

$$\begin{aligned} p(X_i^{au} = 1 | X_j^{au} = 1) > p(X_i^{au} = 0 | X_j^{au} = 1) \\ p(X_i^{au} = 1 | X_j^{au} = 1) > p(X_i^{au} = 1 | X_j^{au} = 0) \end{aligned} \tag{6}$$

Similarly, we have the pairwise dependencies for negatively correlated AU pairs.

$$\begin{aligned} p(X_i^{au} = 1 | X_j^{au} = 1) < p(X_i^{au} = 0 | X_j^{au} = 1) \\ p(X_i^{au} = 1 | X_j^{au} = 1) < p(X_i^{au} = 1 | X_j^{au} = 0) \end{aligned} \tag{7}$$

B Statistic Information of probability constraints

With 6 expressions and 8 AUs, we totally collected 113 probability constraints. Detailed statistics are in Table 1.

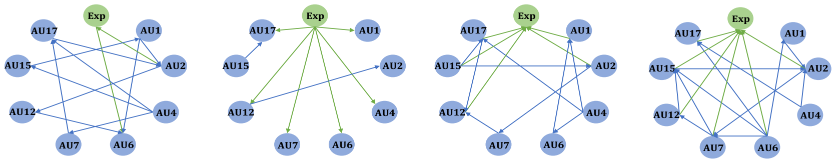
Table 1: Statistics for different types of constraints

Constraints	Strictly Inequality	Inequality	Equality	Total
EI-JAU	50	-	-	50
ED-SAU	30	10	7	47
ED-JAU	16	-	-	16
All the constraints	96	10	7	113

C BN construction with Generic Knowledge

To better understand the effectiveness of probability constraints derived from generic knowledge in BN learning, we compare the BN learned with all the constraints to the BNs learned with a certain type of constraints. We learn three BNs by considering each type of constraints independently, and we then learn a BN by considering all the constraints. We visualize the learned structures[1] in Fig.1. From the Fig.1, we can see that the structures learned with different types of constraints are different. For example, with only expression-independent joint AU constraints(Fig.1(a)), discovered dependencies are mainly among AUs. On the other hand, with only expression-dependent single AU constraints(Fig.1(b)), most of the learned probability dependencies are between expressions and AUs, and dependencies among AUs are rarely captured. Given different types of constraints, we will obtain different BNs capturing different types of probability dependencies. The BN learned with all the constraints are most effective in AU detection and in our experiments, we apply the generic BN(gBN) learned with all the constraints.

[1] Threshold is required to obtain the structure given the learned weighted adjacency matrix. We follow [54] and set the threshold to be 0.3.



(a) EI-JAU (b) ED-SAU (c) ED-JAU (d) All the constraints

Figure 1: Structures of BNs learned with different types of constraints. (a)with only expression independent joint AU probabilities(EI-JAU); (b)with only expression dependent single AU probabilities(ED-SAU); (c)with only expression dependent joint AU probabilities(ED-JAU); (d)with all the probabilities mentioned in (a)-(c)

We then evaluate BNs learned with different types of constraints by considering the corresponding AU detection performance on BP4D, CK+ and MMI. We train AU detection models following Eq.7 in the paper. Results are shown in Table 2. With only one type of constraints, the learned BN captures less informative knowledge compared the the BN learned with all the constraints, and hence produces worse AU detection performance. On all three datasets, the BN learned with all the constraints produces the best AU detection performance. Results in Table 2 further demonstrate the effectiveness of the probability constraints in BN learning. In our experiments for AU detection and FER, we apply the generic BN learned with all the constraints.

Table 2: Evaluation of generic BNs learned with different types of constraints.

Constraints	EI-JAU	ED-SAU	ED-JAU	All the constraints
BP4D	.48	.32	.46	**.56**
CK+	.34	.62	.55	**.69**
MMI	.27	.42	.39	**.47**

D AU-based FER model

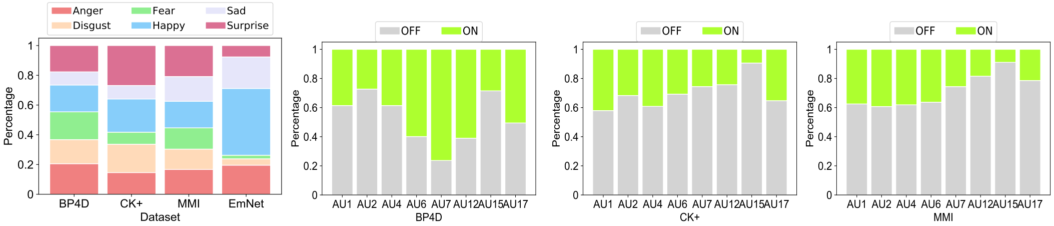
Instead of applying the knowledge model for FER prediction through the Bayes rule, we introduce the trainable AU-based FER model. To compare the FER performance of the knowledge model and AU-based FER model, we first collect the output of the corresponding AUD-BN model as the input of both the knowledge model and the AU-based FER model. We train the AU-based FER model following Eq. 9 in the paper. For the knowledge model, we apply Bayes rule to obtain the expression prediction. The results are shown in Table 3. As we can see from the results, AU-based FER model produces better FER performance than the knowledge model. Though the knowledge model is effective in AU detection, it is not competitive in FER compared to a trainable AU-based FER model. Thus, the AU-based FER model is important in our proposed training framework.

Table 3: Evaluation of the AU-based FER model

Model	BP4D	CK+	MMI
Knowledge Model	21.00	19.11	16.65
AU-based FER	**28.84**	**79.69**	**40.01**

E Information of datasets

We evaluate our proposed methods on four benchmark datasets: BP4D-Spontaneous database, Extended CohnKanande(CK+) database, M&M Initiative(MMI) facial expression database and EmotioNet dataset. For BP4D, CK+, and MMI, annotations for both AUs and 6 basic expressions are collected. For EmotioNet, only 6 basic expressions are collected. Statistical information about the datasets regarding to the expressions and AUs is summarized below.



(a) Expressions (b) AUs in BP4D (c) AUs in CK+ (d) AUs in MMI

Figure 2: Statistics of expressions and AUs in BP4D, CK+, MMI and EmotioNet(EmNet)

F Effects of pre-training

The overall performance of both FER-I and FER-IK without pre-training is worse than the performance with pre-training.

Table 4: Pre-training of FER

Pre-train	model	BP4D	CK+	MMI	EmotioNet
Yes	FER-I	61.68	94.29	67.35	80.85
	FER-IK	83.82	97.59	84.90	95.55
No	FER-I	57.01	79.66	59.64	72.83
	FER-IK	79.76	91.70	82.40	95.50

G Analyse of λ_1 and λ_2

In Table 5, we report the model performance with different values of λ_1 and λ_2 from $\{0.0005, 0.001, 0.005, 0.01, 0.5, 1\}$. On MMI and EmotioNet, the larger values of λ_1 and λ_2 achieve better or comparable performance. While on BP4D and CK+, smaller values of λ_1 and λ_2 produce better performance, in particular on CK+. From the results, we can see that prior knowledge is more important for unbalanced datasets like MMI and noisy datasets like EmotioNet.

Table 5: Performance with different λ_1 and λ_2

model	dataset	0.0005	0.001	0.005	0.01	0.5	1
AUD-EA	BP4D	56.8	56.8	57.5	57.5	57.0	57.0
	CK+	74.4	74.4	74.4	74.3	71.3	71.0
(λ_1)	MMI+	57.0	57.4	57.7	57.7	57.7	57.7
FER-IK	BP4D	83.6	83.8	83.7	83.6	83.6	83.6
	CK+	97.6	97.6	97.6	96.4	94.4	94.5
(λ_2)	MMI+	84.2	84.9	84.9	84.9	84.9	84.9
	EmotioNet	95.3	95.6	95.6	95.6	95.6	95.6