R1 **1.1 The novelty of using generic knowledge.** The existing works except for [51] only consider dataset-specific knowledge, e.g.,[3,22,30,52], and they hence can't generalize well to other datasets. In addition, unlike [51], we propose a novel constraint optimization method to encode the generic knowledge into a BN without requiring any training data.

1.2 The novelty of BN learning. We introduce a constraint optimization method that learns the structure and parameters of a BN simultaneously, while [b] only performs parameter learning. More importantly, our learning is based on only probability constraints without any training data, while [b] requires both training data and constraints.

1.3 Additional source of prior knowledge. The studies([5,6,8]) provide relationships among AUs and expressions and have been used extensively by existing works. [a,c] consider contextual factors that affect emotion perception. They both don't provide the generic knowledge about expression-AUs relationships, and hence they are not considered.

1.4 Justification of using 8 AUs. The 8 AUs are widely considered by existing AU detection models and appear in most benchmark datasets. Our proposed approach can be applied to other AUs as well.

1.5 Clarification on inconsistent comparisons to SOTA. In Tab.6, LP-SM also considers apex frames on CK+, and we apply LP-SM to our BP4D dataset. The comparison to LP-SM is consistent. The reviewer is correct that the comparison with TCAE may be inconsistent. In Tab.8, we apply FMPN-FER and DeepEmotion to our pre-processed datasets for consistent comparisons. STM-Explet, DTAGN and DAM-CNN, though sequence-based methods, are compared with the frame-based methods in the literature and we thus include them. The remaining models apply apex frames and thus the comparisons are consistent.

1.6 Experimental Settings. For the EmotioNet, due to occlusion, AU annotations are usually incomplete and we only perform facial expression recognition evaluation on EmotioNet. [e]('Optimization data') mentioned by the reviewer is for AU recognition in the wild. We will consider a pre-trained VGGFace model in our further work. For the influence of λ_1 and λ_2 on the performance, please refer to the section **5** (at the bottom of the page) for details.

R2 **2.1 The novelty compared to prior work.** Firstly, all of three works mentioned by the reviewer focus on learning a graphical model from a specific dataset to capture probabilistic relationships among different expression-related factors. In contrast, we learn the graphical model from generic and data-independent knowledge and our model can generalize well to different datasets. Secondly, we propose a novel constraint optimization method to learn a BN from probabilistic constraints without requiring any training data, while the graphical models in the three works are either learned using the traditional learning methods(the first paper) or manually specified(the second and third paper). Thirdly, our learned BN is integrated into a deep learning framework for both FER and AU detection, while the three works perform inference on the learned graphical models for predictions. In addition, the first and second paper focus only on one single task.

2.2 Descriptions of expressions and AUs. We agree with the reviewer and there is only one level of facial movement, i.e., the local muscle contraction(AU). Facial expression can be a group of AUs. While AU annotations have less personal bias, they are noisy because the motion is subtle and difficult to distinguish even for experts. Expression is global and easier to label but it is subject to personal bias. In this paper, we don't consider the culture and other contextual factors that can affect the emotion perception and we assume both AU and expression annotations are correct.

2.3 Broader Impact. We agree that facial expression recognition has potential privacy concerns. We will incorporate the examples mentioned by the reviewer into our discussions and will address the privacy concerns.

R3 **3.1 Learning of rBN.** The learning of rBN requires no training data, i.e., no expression and AU labels from a specific dataset. The rBN is learned from probability constraints derived from the generic knowledge via the proposed constraint optimization approach. The same rBN is applied to different datasets in our experiments.

3.2 Effects of each component. We perform ablation studies for each component as shown in Tab.4,5,7, together with additional analysis in supplementary materials. Only the image-based FER model requires pre-training, and we apply a VGG-19 pre-trained on FER2013. In addition, we consider the performance without pre-training as shown in Table 1. Without pre-training, FER-IK can still achieve significant improvement compared to FER-I. The overall performance of both FER-I and FER-IK without pre-training is worse than the performance with pre-training.

3.3 Reproducibility. We will provide the training codes.

3.4 Analysis of λ_1, λ_2. Please refer to the section **5** below.

R4 Thanks for the reviews and we will correct grammatic errors.

5. Analysis of λ_1,λ_2. In Table 2, 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 1: 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

Table 2: 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
	MMI+	84.2	84.9	84.9	84.9	84.9	84.9
(λ_2)	EmotioNet	95.3	95.6	95.6	95.6	95.6	95.6