CROSS-VAE: TOWARDS DISENTANGLING EXPRESSION FROM IDENTITY FOR HUMAN FACES

Haozhe Wu[1,2,3], *Jia Jia*[1,2,3,*], *Lingxi Xie*[4], *Guojun Qi*[5], *Yuanchun Shi*[1,2,3], *Qi Tian*[4]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]Beijing National Research Center for Information Science and Technology (BNRist)
[3]The Institute for Artificial Intelligence, Tsinghua University
[4]Huawei Noah's Ark Lab
[5]Hangzhou Dianzi University
wuhz19@mails.tsinghua.edu.cn, jjia@tsinghua.edu.cn

ABSTRACT

Facial expression and identity are two independent yet intertwined components for representing a face. For facial expression recognition, identity can contaminate the training procedure by providing tangled but irrelevant information. In this paper, we propose to learn clearly disentangled and discriminative features that are invariant of identities for expression recognition. However, such disentanglement normally requires annotations of both expression and identity on one large dataset, which is often unavailable. Our solution is to extend conditional VAE to a crossed version named Cross-VAE, which is able to use partially labeled data to disentangle expression from identity. We emphasis the following novel characteristics of our Cross-VAE: (1) It is based on an independent assumption that the two latent representations' distributions are orthogonal. This ensures both encoded representations to be disentangled and expressive. (2) It utilizes a symmetric training procedure where the output of each encoder is fed as the condition of the other. Thus two partially labeled sets can be jointly used. Extensive experiments show that our proposed method is capable of encoding expressive and disentangled features for facial expression. Compared with the baseline methods, our model shows an improvement of 3.56% on average in terms of accuracy.

Index Terms— Facial expression recognition, Disentangle, Variational Autoencoder

1. INTRODUCTION

Facial expression recognition (FER) is a fundamental but challenging problem in computer vision. In each face image, facial expression and identity (ID), as two orthogonal properties, are entangled together. Due to the intertwined nature of two properties, ID could be regarded as "noise" when we try to explicitly recognize the expression. Thus disentangling ID from expression could be beneficial for the FER task because of cleaner representation. However, such disentanglement usually requires annotations of both expression and identity. Although there are large datasets available for ID [1] and expression [2] recognition respectively, datasets providing both types of annotations [3, 4] are too small for deep neural networks. There are some FER methods [5, 6] that leverage the annotations of both expression and identity to boost FER performance. Those methods can hardly scale to large scale setting due to the requirement of both annotations. To address this issue, we propose a uniformed approach to incorporate the two types of partially labeled data together, and improve FER performance by disentangling facial expression and identity representations from each other.

Researchers have been researching on disentangled representations over the years. However, most of them focus on learning disentangled features for conditional generative tasks [7, 8, 9, 10, 11, 12, 13], few methods [14, 15, 16] have aimed at recognition tasks. Prior disentanglement methods usually distill one attribute from noisy factors. We, on the other hand, explicitly disentangle two orthogonal properties (expression and identity) from each other to learn two discriminative representations.

The major contribution of this work is to generalize conditional VAE (CVAE) [17] into Cross-VAE. The proposed model is a bayesian graphical model which is able to use partially labeled data to disentangle expression and identity from each other. We emphasis the following novelty of our proposed Cross-VAE: (1) It is based on the assumption that two latent representations encoded by two encoders follow two mutually independent multi-variate Gaussian distributions, and each representation takes charge of one property. This ensures both encoded representations to be disentangled and expressive. (2) It takes the advantage of two partially labeled sets, each of which has only ID or expression annotation, then we uniformly train them in our model by providing the output of one encoder to another for image recovery.

[*] Corresponding author

With our proposed Cross-VAE, expressions and identities can thus be disentangled from each other using a mixture of partially labeled data. We perform facial expression recognition experiments on CK+ [4], Oulu [3] and RAF-DB [2] datasets for evaluation. Compared with the baseline methods, Cross-VAE shows an improvement of 3.56% on average in terms of accuracy.

2. APPROACH

For each face image, we have two orthogonal properties, identity and expression, both properties heavily impact image appearance, thus modeling expressions with large identity variations can deliver inaccurate supervision to the model. To avoid noisy training, disentangling one property from the other enables the classifier to classify on cleaner features. Therefore we design the Cross-VAE model to disentangle identity and expression from each other.

In this section, we firstly introduce the Conditional VAE (CVAE) model, which is closely related to our Cross-VAE. Next, we illustrate the formulation of Cross-VAE and explain how it can disentangle two orthogonal latent factors.

2.1. Preliminary: Conditional VAE

Conditional VAE (CVAE) [17] is a directed graphical model that has two variables determining the output variable $\mathbf{x}$. One is the latent variable $\mathbf{z}$ and the other is the input variable $\mathbf{y}$. In general, in this model, parameter estimation is challenging due to intractable posterior inference. Thus, CVAE maximizes the conditional log-likelihood by applying the Stochastic Gradient Variational Bayes (SGVB) [18] framework, where we minimize the evidence lower bound (ELBO) of the log likelihood function. The ELBO is formulated as follows:

$$\begin{aligned}\log p_\theta(\mathbf{x}|\mathbf{y}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}[\log p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z})] \\ - \boldsymbol{KL}[q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})||p(\mathbf{z}|\mathbf{y})] = -\mathcal{L}(\mathbf{x},\mathbf{y};\theta,\phi),\end{aligned} \tag{1}$$

where $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ is the encoder (recognition network), and $p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z})$ is the deocder (generation network). If we want to disentangle $\mathbf{z}$ from the label $\mathbf{y}$, we would make the joint distribution to be independent, *i.e.*, $p_\theta(\mathbf{z}|\mathbf{y}) = p(\mathbf{z})$. Under certain settings, $\mathbf{x}$ is an image, and $\mathbf{y}$ is the label of $\mathbf{x}$, we stipulate $\mathbf{y} \sim \mathrm{cat}(\mathbf{y}|\pi_\theta)$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, where $\mathrm{cat}(\cdot)$ denotes categorical distribution, and $\mathcal{N}(\cdot)$ indicates the multi-variate Gaussian distribution. When optimizing the ELBO, for the convenience of calculating gradients, we use the re-parameterization trick, which implies $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y}) = g_\phi(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon})$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$.

2.2. Formulation: Cross-VAE

Inspired by CVAE which can disentangle the latent variable $\mathbf{z}$ from the input variable $\mathbf{y}$, we propose the Cross-VAE framework. Cross-VAE is also a directed graphical model, which has one output variable $\mathbf{x}$ and two independent latent variables $\mathbf{z}_{\mathbf{emo}}$ and $\mathbf{z}_{\mathbf{id}}$, respectively containing expression information and identity information, and both of them follow a prior distribution of $\mathcal{N}(\mathbf{0},\mathbf{I})$.

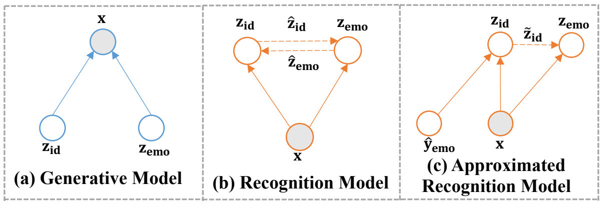
x
$\mathbf{z}_{\mathbf{id}}$
$\mathbf{z}_{\mathbf{emo}}$
(a) Generative Model
$\mathbf{z}_{\mathbf{id}}$
$\hat{\mathbf{z}}_{\mathbf{id}}$
$\mathbf{z}_{\mathbf{emo}}$
$\hat{\mathbf{z}}_{\mathbf{emo}}$
x
(b) Recognition Model
$\mathbf{z}_{\mathbf{id}}$
$\tilde{\mathbf{z}}_{\mathbf{id}}$
$\mathbf{z}_{\mathbf{emo}}$
$\hat{\mathbf{y}}_{\mathbf{emo}}$
x
(c) Approximated Recognition Model

Fig. 1: Graphical representations of Cross-VAE in (a) the generative process; (b) the recognition process in which the dashed lines indicate the $\arg\max$ operation; and (c) the approximated recognition process which deals with the circular dependency problem.

In the Cross-VAE model, $\mathbf{z}_{\mathbf{emo}}$ and $\mathbf{z}_{\mathbf{id}}$ determine the generation of $\mathbf{x}$, as shown in the generative process (decoder) of the model in Fig. 1. We formulate the generative process as $p_\theta(\mathbf{x}|\mathbf{z}_{\mathbf{id}},\mathbf{z}_{\mathbf{emo}})$, where $p_\theta(\cdot)$ measures the probability of generating $\mathbf{x}$ given $\mathbf{z}_{\mathbf{emo}}$ and $\mathbf{z}_{\mathbf{id}}$.

As the reverse of generative process, the recognition process (encoder) would make inference on $\mathbf{z}_{\mathbf{emo}}$ and $\mathbf{z}_{\mathbf{id}}$ given $\mathbf{x}$. During the recognition process, we would estimate $\mathbf{z}_{\mathbf{id}}$ disentangled from expression information, *i.e.*, $p_\theta(\mathbf{z}_{\mathbf{id}}|\mathbf{x},\hat{\mathbf{z}}_{\mathbf{emo}})$, as well as $\mathbf{z}_{\mathbf{emo}}$ disentangled from identity information, *i.e.*, $p_\theta(\mathbf{z}_{\mathbf{emo}}|\mathbf{x},\hat{\mathbf{z}}_{\mathbf{id}})$, the definitions of $\hat{\mathbf{z}}_{\mathbf{emo}}$ and $\hat{\mathbf{z}}_{\mathbf{id}}$ are given in Eqn (2), and the recognition process (encoder) is shown in Fig. 1.

$$\begin{aligned}\hat{\mathbf{z}}_{\mathbf{emo}} &= \mathrm{argmax}\, p_\theta(\mathbf{z}_{\mathbf{emo}}|\mathbf{x},\hat{\mathbf{z}}_{\mathbf{id}}), \\ \hat{\mathbf{z}}_{\mathbf{id}} &= \mathrm{argmax}\, p_\theta(\mathbf{z}_{\mathbf{id}}|\mathbf{x},\hat{\mathbf{z}}_{\mathbf{emo}}).\end{aligned} \tag{2}$$

Again, directly estimating $p_\theta(\mathbf{z}_{\mathbf{id}}|\mathbf{x},\hat{\mathbf{z}}_{\mathbf{emo}})$ and $p_\theta(\mathbf{z}_{\mathbf{emo}}|\mathbf{x},\hat{\mathbf{z}}_{\mathbf{id}})$ is intractable, and we apply the SGVB framework [18] to train the model, in which the evidence lower bound (ELBO) is given as:

$$\begin{aligned}\log p_\theta(\mathbf{x},\hat{\mathbf{z}}_{\mathbf{id}}) \geq \mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{emo}}|\mathbf{x},\hat{\mathbf{z}}_{\mathbf{id}})}[\log p_\theta(\mathbf{x}|\hat{\mathbf{z}}_{\mathbf{id}},\mathbf{z}_{\mathbf{emo}}) \\ + \log p(\hat{\mathbf{z}}_{\mathbf{id}}) + \log p(\mathbf{z}_{\mathbf{emo}}) - \log q_\phi(\mathbf{z}_{\mathbf{emo}}|\mathbf{x},\hat{\mathbf{z}}_{\mathbf{id}})],\end{aligned} \tag{3}$$

$$\begin{aligned}\log p_\theta(\mathbf{x},\hat{\mathbf{z}}_{\mathbf{emo}}) \geq \mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{id}}|\mathbf{x},\hat{\mathbf{z}}_{\mathbf{emo}})}[\log p_\theta(\mathbf{x}|\mathbf{z}_{\mathbf{id}},\hat{\mathbf{z}}_{\mathbf{emo}}) \\ + \log p(\hat{\mathbf{z}}_{\mathbf{emo}}) + \log p(\mathbf{z}_{\mathbf{id}}) - \log q_\phi(\mathbf{z}_{\mathbf{id}}|\mathbf{x},\hat{\mathbf{z}}_{\mathbf{emo}})].\end{aligned} \tag{4}$$

However, optimizing Eqns (3) and (4) has circular dependency, *i.e.* $\hat{\mathbf{z}}_{\mathbf{id}}$ and $\hat{\mathbf{z}}_{\mathbf{emo}}$ depend on each other and thus we are unable to estimate them. We perform an approximation which substitutes $\hat{\mathbf{z}}_{\mathbf{id}}$ and $\hat{\mathbf{z}}_{\mathbf{emo}}$ by $\tilde{\mathbf{z}}_{\mathbf{id}}$ and $\tilde{\mathbf{z}}_{\mathbf{emo}}$, respectively. This provides a possible manner of optimizing the evidence lower bound, which works as follows. First, we approximate $\hat{\mathbf{z}}_{\mathbf{id}}$ by using a pretrained expression classifier $p(\mathbf{y}_{\mathbf{emo}}|\mathbf{x})$:

$$\begin{aligned}\hat{\mathbf{z}}_{\mathbf{id}} &= \mathrm{argmax}\, p_\theta(\mathbf{z}_{\mathbf{id}}|\mathbf{x},\hat{\mathbf{z}}_{\mathbf{emo}}) \\ &\approx \mathrm{argmax}\, q_\phi(\mathbf{z}_{\mathbf{id}}|\mathbf{x},\hat{\mathbf{z}}_{\mathbf{emo}}) \\ &\approx \mathrm{argmax}\, q_\phi(\mathbf{z}_{\mathbf{id}}|\mathbf{x},\hat{\mathbf{y}}_{\mathbf{emo}}) = \tilde{\mathbf{z}}_{\mathbf{id}},\end{aligned} \tag{5}$$

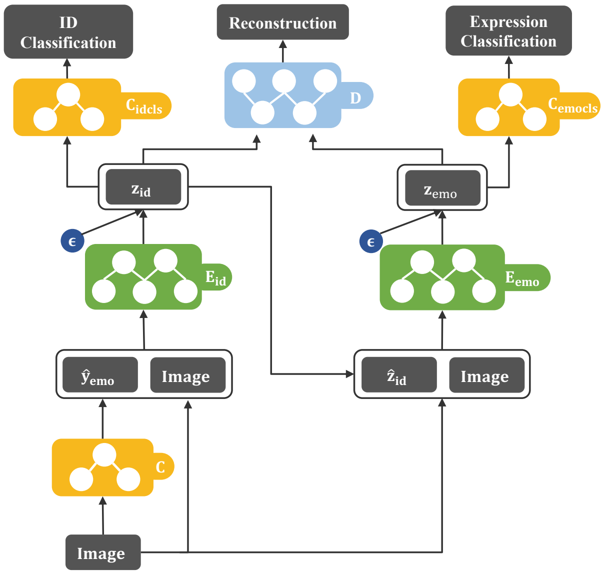

Fig. 2: The overall framework of Cross-VAE

where $\hat{\mathbf{y}}_{\mathbf{emo}} = \operatorname{argmax} p(\mathbf{y}_{\mathbf{emo}}|\mathbf{x})$. Next, we approximate $q_\phi(\mathbf{z}_{\mathbf{id}}|\mathbf{x}, \hat{\mathbf{z}}_{\mathbf{emo}})$ in Eqn (4) by $q_\phi(\mathbf{z}_{\mathbf{id}}|\mathbf{x}, \hat{\mathbf{y}}_{\mathbf{emo}})$. Finally, $\hat{\mathbf{z}}_{\mathbf{emo}}$ is estimated using

$$\begin{aligned}\hat{\mathbf{z}}_{\mathbf{emo}} &= \operatorname{argmax} p_\theta(\mathbf{z}_{\mathbf{emo}}|\mathbf{x}, \hat{\mathbf{z}}_{\mathbf{id}}) \\ &\approx \operatorname{argmax} q_\phi(\mathbf{z}_{\mathbf{emo}}|\mathbf{x}, \tilde{\mathbf{z}}_{\mathbf{id}}) = \tilde{\mathbf{z}}_{\mathbf{emo}}.\end{aligned} \tag{6}$$

Following the approximations above, the approximated recognition process is shown in Fig. 1, we can rewrite the evidence lower bound (ELBO) as follows:

$$\begin{aligned}\mathcal{L}_{\text{elbo}} = &-\mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{id}}|\mathbf{x}, \hat{\mathbf{y}}_{\mathbf{emo}})}[\log p_\theta(\mathbf{x}|\mathbf{z}_{\mathbf{id}}, \tilde{\mathbf{z}}_{\mathbf{emo}})] \\ &+ \boldsymbol{KL}[q_\phi(\mathbf{z}_{\mathbf{id}}|\mathbf{x}, \hat{\mathbf{y}}_{\mathbf{emo}})||p(\mathbf{z}_{\mathbf{id}})] \\ &- \mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{emo}}|\mathbf{x}, \tilde{\mathbf{z}}_{\mathbf{id}})}[\log p_\theta(\mathbf{x}|\tilde{\mathbf{z}}_{\mathbf{id}}, \mathbf{z}_{\mathbf{emo}})] \\ &+ \boldsymbol{KL}[q_\phi(\mathbf{z}_{\mathbf{emo}}|\mathbf{x}, \tilde{\mathbf{z}}_{\mathbf{id}})||p(\mathbf{z}_{\mathbf{emo}})],\end{aligned} \tag{7}$$

where $\tilde{\mathbf{z}}_{\mathbf{id}}$, $\tilde{\mathbf{z}}_{\mathbf{emo}}$ and $\hat{\mathbf{y}}_{\mathbf{emo}}$ do not have circular dependency.

In our framework, $p_\theta(\mathbf{x}|\mathbf{z}_{\mathbf{id}}, \tilde{\mathbf{z}}_{\mathbf{emo}})$ and $p_\theta(\mathbf{x}|\tilde{\mathbf{z}}_{\mathbf{id}}, \mathbf{z}_{\mathbf{emo}})$ share the same decoder, while encoder $q_\phi(\mathbf{z}_{\mathbf{id}}|\mathbf{x}, \hat{\mathbf{y}}_{\mathbf{emo}})$ and encoder $q_\phi(\mathbf{z}_{\mathbf{emo}}|\mathbf{x}, \tilde{\mathbf{z}}_{\mathbf{id}})$ don't share parameters, we detailedly illustrate the overall framework of Cross-VAE in Fig 2. In order to enhance the disentangle effect, we multiply the KL loss term with a weight of β, like what was done by β-VAE [8]. We add two classifiers C_{emocls} and C_{idcls} respectively on two latent variables so as to supervise the encoder to learn discriminative representations, and denote the objective of two classifiers as $\mathcal{L}_{\text{cls}}$.

In summary, the objective of Cross-VAE is written as:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{cls}} + \alpha \cdot \mathcal{L}_{\text{elbo}}. \tag{8}$$

3. EXPERIMENTS

In this section, we show extensive experimental results of our model. We conduct experiments on three public datasets: Oulu-CASIA [3], CK+ [4], and RAF-DB datasets [2] for facial expression recognition. Our Cross-VAE model shows better performance compared to baseline methods. In addition, we conduct qualitative experiments to show that our model can disentangle ID and expression into two independent representations.

3.1. Dataset Descriptions

Oulu-CASIA. The Oulu-CASIA dataset [3] contains 2,880 image sequences collected from 80 subjects and six basic expressions. During training and testing phases, we follow the previous method [19] to select the last three frames from 480 sequences to conduct a 10-fold cross validation protocol.

CK+. CK+ [4] is an extensively used dataset for FER. It contains 327 labeled sequences from 118 subjects consisting of seven expressions. Following the previous setting [5], we choose the last three frames of each sequence to conduct a 8-fold cross validation protocol.

RAF-DB. RAF-DB [2] is a real-world wild dataset downloaded from the Internet. It contains 12,271 training images and 3,068 test images. Images are manually labeled with seven kinds of expressions.

For all datasets above, we use standard pre-processing. We detect facial landmarks in each image by MTCNN [20], then crop and align facial images using similarity transformation. When MTCNN does not detect any face in the image, we retain it if the image is in test set, otherwise it is simply discarded. We resize each image to 100 × 116 and normalize the intensity values from $[0, 255]$ to $[-1, 1]$. In order to alleviate over-fitting, several data augmentation techniques are applied, including random horizontal flipping and random image crop into 96 × 112 pixels. During testing, each image is center-cropped into a 96 × 112 scale and sent into prediction.

3.2. Implementation Details

Cross-VAE model has three modules, expression encoder, identity encoder and decoder. Both expression encoder and identity encoder use 18-layer resnet [21] in CNN architecture. For decoder, firstly we use a FC layer to fuse $\mathbf{z}_{\mathbf{id}}$ and $\mathbf{z}_{\mathbf{emo}}$, then we use a five-layer deconvolution network to decode the latent representation. For classifier, only one FC layer is used. As for the weight in $\mathcal{L}_{\text{all}}$, we set $\alpha = 1 \times 10^{-4}$, $\beta = 5$ in all settings. During the training of Cross-VAE, we use adam optimizer [22] with 1×10^{-4} learning rate and 5×10^{-4} weight decay.

Method	CK+	Oulu	RAF-DB
3D-CNN-DAP [24]	92.40	-	-
STM-ExpLet [25]	94.19	74.59	-
IACNN [5]	95.37	-	-
(N+M) Softmax [6]	97.1	-	-
DTAGN [26]	97.25	81.46	-
DeRL [27]	97.30	88.00	-
GCNet [28]	97.93	86.30	-
AdaLBP [3]	-	73.54	-
Atlases [29]	-	75.52	-
FN2EN [19]	-	87.71	-
Center Loss [2]	-	-	82.86
PG-CNN [30]	-	-	83.27
PAT-Resnet [31]	-	-	84.19
Resnet-18	86.52	86.00	83.28
Cross-VAE	**94.96**	**86.87**	**84.81**

Table 1: Accuracy (%) on the CK+, Oulu-CASIA and RAF-DB datasets.

3.3. Facial Expression Recognition Results

In this section, we compare our method with several baseline methods on FER task. The expression encoder in our model has same backbone as the Resnet-18, thus we compare our method with Resnet-18 in particular. For RAF-DB, we initialize parameters of Resnet-18 pretrained by Imagenet [23]. For Oulu-CASIA and CK+, we initialize parameters pretrained by RAF-DB.

CK+. Table 1 reports the average accuracy of a 8-fold cross validation. The backbone of our expression encoder is same as Resnet-18, while our model achieves an accuracy improvement of 8.43% compared to Resnet-18, which shows the validness of disentangled feature. And our model shows close performance to IACNN [5], which requires both subject and expression label of each image.

Oulu-CASIA. Table 1 reports the average accuracy of a 10-fold cross validation. Cross-VAE shows an accuracy improvement of 0.87% compared to the Resnet-18. Next, we compare our method with DTAGN [26], a jointly fine tune method which uses sequences of facial landmarks and facial images to conduct expression recognition. Our method performs 5.4% better than DTGAN, showing the effect of disentangled feature.

RAF-DB. RAF-DB is a wild dataset with larger training and testing size. Our Cross-VAE model shows an accuracy improvement of 1.53% compared to resnet-18, and achieves highest accuracy compared with those state-of-the-art methods. Previous methods generally can't utilze identity information in such wild dataset, while our method can use large ID annotated dataset to improve performance on wild expression dataset. The satisfactory performance on the wild dataset shows the effectiveness of disentangled expression representations.

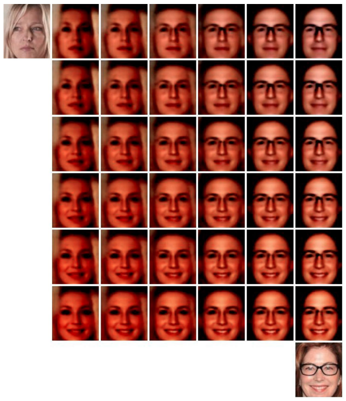

Fig. 3: Visulization of disentanglement.

To summarize, we use ID annotated dataset to learn expression representation disentangled from identity, improving performance on CK+, Oulu-CASIA and RAF-DB.

3.4. Effect of Disentangle

In this section, we qualitatively analyze the disentangle effect of Cross-VAE. We sample an image pair from CASIA dataset $\mathbf{x_1}, \mathbf{x_2}$ and use Cross-VAE to extract $\mathbf{z_{emo1}}, \mathbf{z_{emo2}}, \mathbf{z_{id1}}, \mathbf{z_{id2}}$. Then we do linear interpolation between $\mathbf{z_{emo1}}$ and $\mathbf{z_{emo2}}$, $\mathbf{z_{id1}}$ and $\mathbf{z_{id2}}$ to generate an image matrix, results are shown in Fig. 3. From the generated images, we can see that Cross-VAE can capture the subtle change of expression and identity and learn disentangled representations.

4. CONCLUSIONS

In this paper, we propose Cross-VAE, a generalized framework beyond CVAE, so as to allow the disentanglement of both expression and identity representations from faces. Towards optimizing the approximated evidence lower bound (ELBO), our Cross-VAE provides a solid explanation of its success on disentangling expressions from identities. With its crossed formulation, our model is able to take the advantage of training two partially labelled dataset simultaneously. Experiments show consistently performance gain for the facial expression recognition task and proves the effectiveness of our disentangle mechanism for encoding two orthogonal discriminative features.

5. ACKNOWLEDGMENTS

Thanks for advice of Hang Zhou from CUHK, his advice improves our work further. This work is supported by National Key Research and Development Plan (2016YFB1001200), Beijing Academy of Artificial Intelligence (BAAI), and National Natural Science Foundation of China (61521002, 61831022).

6. REFERENCES

[1] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[2] Shan Li, Weihong Deng, and JunPing Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2852–2861.

[3] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[4] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain 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*. IEEE, 2010, pp. 94–101.

[5] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong, "Identity-aware convolutional neural network for facial expression recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 558–565.

[6] Xiaofeng Liu, BVK 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, pp. 20–29.

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in neural information processing systems*, 2016, pp. 2172–2180.

[8] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017.

[9] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum, "Deep convolutional inverse graphics network," in *Advances in neural information processing systems*, 2015, pp. 2539–2547.

[10] Siddharth Narayanaswamy, T Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr, "Learning disentangled representations with semi-supervised deep generative models," in *Advances in Neural Information Processing Systems*, 2017, pp. 5925–5935.

[11] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations," *arXiv preprint arXiv:1711.00848*, 2017.

[12] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 9299–9306.

[13] Hyunjik Kim and Andriy Mnih, "Disentangling by factorising," *arXiv preprint arXiv:1802.05983*, 2018.

[14] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al., "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," in *Advances in Neural Information Processing Systems*, 2018, pp. 1230–1241.

[15] Luan Tran, Xi Yin, and Xiaoming Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1415–1424.

[16] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang, "Exploring disentangled feature representation beyond face identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2080–2089.

[17] Kihyuk Sohn, Honglak Lee, and Xinchen Yan, "Learning structured output representation using deep conditional generative models," in *Advances in neural information processing systems*, 2015, pp. 3483–3491.

[18] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[19] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 118–126.

[20] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," 2009.

[24] Mengyi Liu, Shaoxin Li, Shiguang Shan, Ruiping Wang, and Xilin Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Asian conference on computer vision*. Springer, 2014, pp. 143–157.

[25] 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*, 2014, pp. 1749–1756.

[26] 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*, 2015, pp. 2983–2991.

[27] 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*, 2018, pp. 2168–2177.

[28] Youngsung Kim, ByungIn Yoo, Youngjun Kwak, Changkyu Choi, and Junmo Kim, "Deep generative-contrastive networks for facial expression recognition," *arXiv preprint arXiv:1703.07140*, 2017.

[29] Yimo Guo, Guoying Zhao, and Matti Pietikäinen, "Dynamic facial expression recognition using longitudinal facial expression atlases," in *Computer Vision–ECCV 2012*, pp. 631–644. Springer, 2012.

[30] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen, "Patch-gated cnn for occlusion-aware facial expression recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2209–2214.

[31] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong, "Probabilistic attribute tree in convolutional neural networks for facial expression recognition," *arXiv preprint arXiv:1812.07067*, 2018.