D^{3}Net: Dual-Branch Disturbance Disentangling Network for Facial Expression Recognition

Rongyun Mo
Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University,
Xiamen, China
morongyun@stu.xmu.edu.cn

Yan Yan*
Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University,
Xiamen, China
yanyan@xmu.edu.cn

Jing-Hao Xue
Department of Statistical Science, University College London,
London, UK
jinghao.xue@ucl.ac.uk

Si Chen
School of Computer and Information Engineering, Xiamen University of Technology,
Xiamen, China
chensi@xmut.edu.cn

Hanzi Wang
Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University,
Xiamen, China
hanzi.wang@xmu.edu.cn

ABSTRACT

One of the main challenges in facial expression recognition (FER) is to address the disturbance caused by various disturbing factors, including common ones (such as identity, pose, and illumination) and potential ones (such as hairstyle, accessory, and occlusion). Recently, a number of FER methods have been developed to explicitly or implicitly alleviate the disturbance involved in facial images. However, these methods either consider only a few common disturbing factors or neglect the prior information of these disturbing factors, thus resulting in inferior recognition performance. In this paper, we propose a novel Dual-branch Disturbance Disentangling Network (D^{3}Net), mainly consisting of an expression branch and a disturbance branch, to perform effective FER. In the disturbance branch, a label-aware sub-branch (LAS) and a label-free sub-branch (LFS) are elaborately designed to cope with different types of disturbing factors. On the one hand, LAS explicitly captures the disturbance due to some common disturbing factors by transfer learning on a pretrained model. On the other hand, LFS implicitly encodes the information of potential disturbing factors in an unsupervised manner. In particular, we introduce an Indian buffet process (IBP) prior to model the distribution of potential disturbing factors in LFS. Moreover, we leverage adversarial training to increase the differences between disturbance features and expression features, thereby enhancing the disentanglement of disturbing factors. By disentangling the disturbance from facial images, we are able to extract discriminative expression features. Extensive experiments demonstrate that our proposed method performs favorably against several state-of-the-art FER methods on both in-the-lab and in-the-wild databases.

CCS CONCEPTS

- **Computing methodologies → Image representations**;

KEYWORDS

Facial expression recognition, Multi-task learning, Transfer learning, Disentangled representation learning

ACM Reference Format:
Rongyun Mo, Yan Yan, Jing-Hao Xue, Si Chen, and Hanzi Wang. 2021. D^{3}Net: Dual-Branch Disturbance Disentangling Network for Facial Expression Recognition. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475249

1 INTRODUCTION

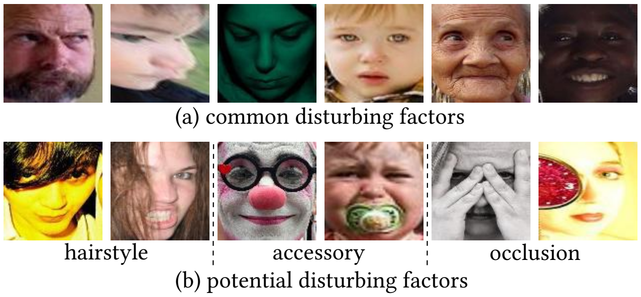
Recently, with the rapid development of deep learning, facial expression recognition (FER) has made remarkable progress in multimedia and computer vision [32, 36, 39, 40, 45], mainly due to its practical significance in human-computer interaction, health care systems, digital entertainment, *etc* [6]. However, FER is still a very challenging problem. This is mostly because facial images usually involve the disturbance caused by various disturbing factors, resulting in large appearance variations. Therefore, it is of great importance to disentangle the disturbance from facial images for effective FER.

In general, facial images are easily affected by some common disturbing factors (e.g., illumination and pose) that are often ubiquitous in FER databases and seriously deteriorate the recognition accuracy. In Figure 1(a), variations of identity, pose, illumination, gender, race, and age exist in facial images and interfere the extraction of expression features. Extensive deep learning-based FER methods [4, 30, 32, 43] have been proposed to exploit the labels of common disturbing factors and explicitly disentangle the disturbance.

Unfortunately, the above methods ignore the fact that there may exist many other disturbing factors. As shown in Figure 1(b), facial images are apparently influenced by some potential disturbing

[*] Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475249


(a) common disturbing factors

(b) potential disturbing factors

Figure 1: Some facial images influenced by (a) common disturbing factors (such as identity, pose, illumination, gender, race, and age) and (b) potential disturbing factors (such as hairstyle, accessory, and occlusion).

factors (such as hairstyle, accessory, and occlusion), which also degrade the FER performance. Note that most FER databases do not provide the label information for potential disturbing factors that can vary in different FER databases. Moreover, it is difficult to predefine all the possible disturbing factors for FER. In recent years, some methods [2, 21, 42] have been developed to implicitly suppress the disturbance in facial images in an unsupervised manner without specifying disturbing factors. Generally, these methods make no difference between common disturbing factors and potential ones.

For FER, it is important to address both common disturbing factors and potential ones. However, previous methods either focus on only a few common disturbing factors [4, 32] or do not fully take into account the prior information of common disturbing factors [2, 42], leading to sub-optimal performance.

To tackle the above problem, we propose a novel Dual-branch Disturbance Disentangling Network (called D^{3}Net) by elaborately performing both explicit and implicit disentanglement of various disturbing factors for FER. D^{3}Net consists of a shared backbone network, and two task-specific branches (i.e., an expression branch and a disturbance branch) that learn expression features and disturbance features, respectively. For the disturbance branch, it contains two sub-branches (i.e., a label-aware disturbance sub-branch (LAS) and a label-free disturbance sub-branch (LFS)) to extract disturbance features for common disturbing factors and potential ones, respectively. By jointly training the two task-specific branches on the shared backbone network, we can effectively disentangle comprehensive disturbance information from the expression information.

The contributions of this paper are summarized as follows:

(1) We propose a novel D^{3}Net method which elaborately designs a disturbance branch to disentangle various disturbing factors in the FER database. In particular, LAS is trained to explicitly capture the information of common disturbing factors by transfer learning, while LFS is developed to implicitly learn the information of potential disturbing factors in a fully unsupervised manner. In this way, our method is able to extract discriminative expression features by suppressing various disturbing factors.

(2) We introduce an Indian buffet process (IBP) prior to model the distribution of potential disturbing factors in LFS. Moreover, we leverage adversarial training to distinguish label-free disturbance features from expression features. Combining the IBP prior with adversarial training is advantageous to learn the latent structure of potential disturbing factors from facial images in an unsupervised manner. To the best of our knowledge, this is the first work to make use of the IBP prior to perform implicit disturbance disentanglement for FER.

(3) We carry out extensive experiments on public FER databases, including three in-the-lab databases and two in-the-wild databases. Experimental results demonstrate that our proposed method achieves superior performance against several state-of-the-art FER methods.

2 RELATED WORK

Explicit Disturbance-Disentangled FER Methods. These methods explicitly disentangle the disturbance by exploiting the label information of common disturbing factors. For example, Wang *et al.* [36] employ an encoder that is adversarially trained with two discriminators to address pose variations and identity bias. Some methods [4, 43, 46] apply generative adversarial network (GAN) [10] to learn identity-invariant or pose-invariant features, by generating facial images with different identity labels or pose labels. Ruan *et al.* [32] propose a deep disturbance-disentangled learning (DDL) method which simultaneously disentangles several common disturbing factors based on adversarial transfer learning.

The above methods address common disturbing factors but overlook the influence caused by potential disturbing factors that also harm the FER performance. In fact, it is difficult to specify all the possible disturbing factors for FER. As a result, it is not a trivial task to disentangle the disturbance in a fully-supervised manner.

Implicit Disturbance-Disentangled FER Methods. These methods implicitly disentangle the disturbance without specifying disturbing factors. Based on adversarial training, Halawa *et al.* [16] develop an unsupervised method to disentangle the disturbance for FER. Some methods [2, 21, 42] explore the differences between the expression component and the neutral component to learn expression features. However, these methods depend highly on neutral facial images. Moreover, they may fail to classify facial images exhibiting weak expressions similar to neutral expressions.

It is widely acknowledged that some common disturbing factors are of great significance for achieving excellent FER performance. The above methods, however, do not fully explore this important prior information for FER, and thus may not well reduce the disturbance due to these disturbing factors. In this paper, we innovatively design two sub-branches to explicitly capture the information of common disturbing factors and implicitly encode the information of potential disturbing factors, respectively. Hence, we can effectively eliminate the disturbance caused by various disturbing factors.

Unsupervised Disentangled Representation Learning (UDRL) Methods. UDRL methods aim to identify the underlying factors from observed data [3], and they are mostly based on the variational auto-encoder (VAE) [23]. Higgins *et al.* [18] develop β-VAE with an isotropic Gaussian prior to enhance the independence of each element of latent variables. Inspired by β-VAE, some Gaussian posterior approximation-based methods [5, 20] are developed to address the problem of β-VAE that it cannot effectively balance the tradeoff between disentanglement and image reconstruction. However, the disentanglement ability of these methods generally degrades when the number of underlying factors increases [15].

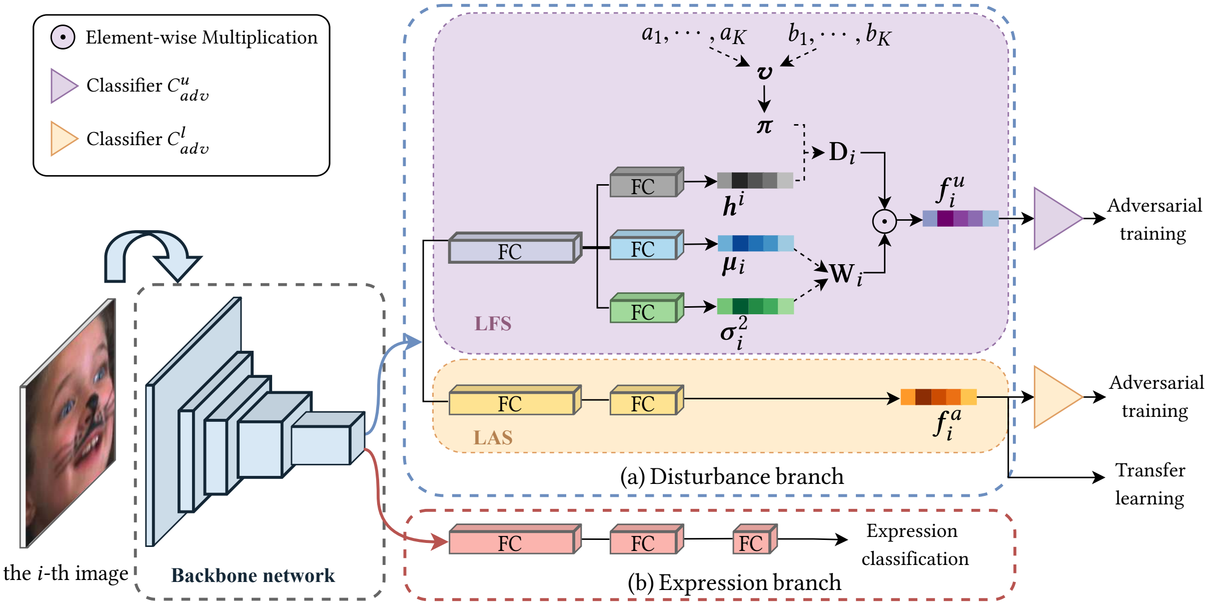


Figure 2: Overview of our proposed D^{3}Net method, containing a shared backbone network to extract the shared features, (a) a disturbance branch consisting of two sub-branches (i.e., label-aware disturbance sub-branch (LAS) and label-free disturbance sub-branch (LFS)) to capture the disturbance information in explicit and implicit ways, respectively, and (b) an expression branch to capture the expression information.

Later, a non-parametric prior is also combined with VAE to perform disentanglement of task-irrelevant factors for skin lesion classification [15].

For FER, the number of all the possible disturbing factors is difficult to be predefined and can be large. Hence, Gaussian posterior approximation-based methods may be inappropriate to perform disturbance disentanglement in FER. In this paper, we capitalize on a non-parametric prior, i.e., the Indian buffet process (IBP) prior [12], which enables to obtain a richer posterior approximation, to implicitly disentangle potential disturbing factors. Note that previous methods [15, 33] that combine the IBP prior with VAE aim to either improve image reconstruction or achieve a tradeoff between disentanglement and image reconstruction. In contrast, considering the characteristics of the FER task, instead of adopting the decoder for image reconstruction, we perform adversarial training between disturbance features and expression features. Such a manner greatly enhances the performance of disturbance disentanglement for FER.

3 PROPOSED METHOD

3.1 Overview

D^{3}Net is comprised of a shared backbone network, an expression branch, and a disturbance branch, as shown in Figure 2. In this paper, we employ the commonly used ResNet-18 [17] as the backbone to extract the shared features, where the expression information is heavily entangled with the disturbance information. Based on the shared backbone network, the expression branch and the disturbance branch are developed to extract expression features and disturbance features, respectively. For the expression branch, we adopt a set of fully-connected (FC) layers to predict facial expressions. For the disturbance branch, two sub-branches, consisting of a label-aware disturbance sub-branch (LAS) and a label-free disturbance sub-branch (LFS), are elaborately designed to capture the disturbance information in different ways. Concretely, LAS and LFS extract the disturbance features of common disturbing factors and potential disturbing factors in explicit and implicit ways, respectively. Finally, by jointly optimizing the expression branch and the disturbance branch on the shared backbone network, our proposed D^{3}Net can fully disentangle the disturbance from facial images, and thus effectively extract discriminative expression features for FER.

More specifically, in LAS, we adopt transfer learning to explicitly capture the label-aware disturbance information with a pretrained model (which is learned to identify multiple common disturbing factors trained on large-scale face databases). This enables LAS to extract discriminative features of these common disturbing factors, even when the labels of disturbing factors are not available in the FER database. Meanwhile, in LFS, we take advantage of UDRL to implicitly encode the label-free disturbance information of potential disturbing factors that are not considered in LAS. In particular, a non-parametric IBP prior is introduced to model the distribution of potential disturbing factors. Moreover, adversarial training is employed to enlarge the differences between disturbance features and expression features. By combining LAS with LFS in an integrated network, we are able to sufficiently capture the disturbance information from facial images.

During the inference stage, the test images are fed into the trained backbone network and expression branch to extract features for classification.

3.2 Expression Branch

Following the backbone network, the expression branch consists of three FC layers. Assume that a training set $\mathcal{D}_t = \{(\mathbf{x}_t^i, y_t^i)\}_{i=1}^N$ contains N facial expression images in the FER database, where $\mathbf{x}_t^i$ and y_t^i represent the i-th facial image and its corresponding expression label, respectively. We train the expression branch by minimizing the cross-entropy loss, which is defined as

$$\mathcal{L}_{exp} = -\sum_{i=1}^{N}\sum_{c=1}^{C} \mathbb{1}_{[c=y_t^i]} \log(\mathcal{P}_e(E_{exp}(\mathbf{x}_t^i))), \tag{1}$$

where C is the number of expression categories; E_{exp} represents the expression feature extractor (the backbone network and the first two FC layers in the expression branch); $\mathcal{P}_e$ is the prediction function (the final FC layer) for classifying expressions; $\mathbb{1}_{[c=y_t^i]}$ equals to 1 when $c = y_t^i$, and 0 otherwise.

3.3 Disturbance Branch

The disturbance branch has two sub-branches, i.e., LAS and LFS.

3.3.1 Label-Aware Disturbance Sub-Branch (LAS). LAS is developed to capture the disturbance information of common disturbing factors. However, only the labels of identity and pose are available in most FER databases. Fortunately, some large-scale face databases offer the labels of common disturbing factors. For example, Multi-PIE [13] and RAF-DB [25] provide the labels of identity, pose, illumination, and those of gender, race, age, respectively. Therefore, we are able to take advantage of transfer learning to exploit these available labels, achieving explicit disentanglement of common disturbing factors in the FER database.

Inspired by DDL [32], we pretrain a model encoding the disturbance information of common disturbing factors in the face databases, thus providing reference disturbance features for training LAS. More specifically, given a face database, its training set $\mathcal{D}_s$ with R images can be denoted as $\mathcal{D}_s = \{(\mathbf{x}_s^i, \mathbf{s}_i)\}_{i=1}^R$, where $\mathbf{x}_s^i$ is the i-th training image and $\mathbf{s}_i = [s_i^1, \cdots, s_i^M]$ denotes the corresponding labels of M disturbing factors. A pretrained model is learned based on the labels of M disturbing factors. For the j-th disturbing factor, its classification loss function $\mathcal{L}_p^j$ is defined as

$$\mathcal{L}_p^j = -\sum_{i=1}^{R}\sum_{c=1}^{C_j} \mathbb{1}_{[c=s_i^j]} \log(\mathcal{P}_j(E_j(\mathbf{x}_s^i))), \tag{2}$$

where C_j is the class number of the j-th disturbing factor; E_j and $\mathcal{P}_j$ represent the feature extractor and the prediction function for classifying the j-th disturbing factor, respectively.

Thus, the pretrained model is trained by optimizing

$$\mathcal{L}_p = \sum_{j=1}^{M} \mathcal{L}_p^j, \tag{3}$$

where $\mathcal{L}_p$ denotes the classification loss of the pretrained model. More details of the pretrained model can be referred in [32].

Based on the above pretrained model, LAS comprising of two FC layers is then trained on the FER database. Given a facial expression image $\mathbf{x}_t^i$ from the FER database, the label-aware disturbance feature extracted by LAS is denoted as $\boldsymbol{f}_i^a \in \mathbb{R}^{1\times D}$, while the reference disturbance feature $\boldsymbol{f}_i^r \in \mathbb{R}^{1\times D}$ extracted from the pretrained model is defined as $\boldsymbol{f}_i^r = \sum_{j=1}^{M'} \boldsymbol{f}_{i,j}^p$. Here, M' denotes the number of selected common disturbing factors for training LAS, and $\boldsymbol{f}_{i,j}^p \in \mathbb{R}^{1\times D}$ is a disturbance feature extracted by E_j. To effectively transfer the knowledge from the pretrained model to LAS, we adopt the Kullback-Liebler (K-L) divergence to constrain the probability distributions of the features extracted by the pretrained model and LAS to be as close as possible. Hence, the loss function for training LAS is defined as

$$\mathcal{L}_{LAS} = \sum_{i=1}^{N} D_{KL}(\boldsymbol{f}_i^r || \boldsymbol{f}_i^a) = \sum_{i=1}^{N}\sum_{j=1}^{D} \bar{f}_{i,j}^r \cdot \log \frac{\bar{f}_{i,j}^r}{\bar{f}_{i,j}^a}, \tag{4}$$

where $D_{KL}(\cdot||\cdot)$ represents the K-L divergence; $\bar{f}_{i,j}^a = \text{softmax}(f_{i,j}^a)$ and $\bar{f}_{i,j}^r = \text{softmax}(f_{i,j}^r)$ (here, softmax$(\cdot)$ indicates the softmax operation); $f_{i,j}^a$ and $f_{i,j}^r$ are the j-th elements of $\boldsymbol{f}_i^a$ and $\boldsymbol{f}_i^r$, respectively.

3.3.2 Label-Free Disturbance Sub-Branch (LFS). LFS is designed to perform implicit disentanglement of potential disturbing factors in an unsupervised way. Note that most representative UDRL methods [5, 18, 20] are based on a Gaussian approximation of the posterior density, where the disentanglement ability can be severely affected by the increasing number of underlying factors [15]. However, the possible disturbing factors are usually unknown, their number can be large, and their presence is often sparse for the FER task. Thus, the Gaussian posterior approximation-based methods may not be appropriate for learning disturbance features in FER.

In this paper, to address the above problem, we introduce an IBP prior, which enables to give a richer posterior approximation, to effectively model the distribution of a number of potential disturbing factors in LFS. As shown in Figure 2, LFS contains one hidden layer (an FC layer) and three output layers (three FC layers), which output the label-free disturbance features.

On the one hand, the IBP (also known as the Beta-Bernoulli process) [11] is constructed as a prior on a sparse binary matrix $\mathbf{D} \in \{0,1\}^{N\times K}$ (which is defined as the disturbance occurrence matrix in this paper), where K represents the number of potential disturbing factors. $\mathbf{D}_i$ denotes the i-th row of $\mathbf{D}$. In $\mathbf{D}$, the element $d_{i,k}$ equals to 1 if the k-th disturbing factor appears in the i-th facial image and 0 otherwise. Each $d_{i,k}$ obeys the Bernoulli distribution (i.e., $p(d_{i,k}) = Bernoulli(\pi_k)$, where π_k represents the occurrence probability of the k-th disturbing factor. In practice, π_k is generated by a stick-breaking method [34]. In this way, π_k can be formulated as the product of a set of independent random variables $\boldsymbol{v} = \{v_1, v_2, \cdots, v_k, \cdots\}$, i.e., $\pi_k = \prod_{j=1}^k v_j$, where each variable v_k follows a Beta distribution (i.e., $p(v_k) = Beta(\alpha, 1)$). Hence, given $\boldsymbol{v}$, the prior density of $\mathbf{D}_i$ is formulated as

$$p(\mathbf{D}_i|\boldsymbol{v}) = \prod_{k=1}^{K} Bernoulli(\pi_k). \tag{5}$$

On the other hand, we define $\mathbf{W} \in \mathbb{R}^{N\times K}$ as a weight matrix [12], whose prior density is assumed to follow a Gaussian distribution. Given a weight vector $\mathbf{W}_i$ (i.e., the i-th row of $\mathbf{W}$), we have $p(\mathbf{W}_i) = \mathcal{N}(0, \mathbf{I}_K)$, where $\mathcal{N}$ denotes the Gaussian function with an identity matrix $\mathbf{I}_K \in \mathbb{R}^{K\times K}$.

Then, the K-L losses between the posterior densities and the prior densities of the variables $\boldsymbol{v}$, the disturbance occurrence matrix $\mathbf{D}$, and the weight matrix $\mathbf{W}$ are jointly minimized to train LFS. Thus, we can perform implicit disentanglement of potential disturbing factors.

Specifically, to enable the training of the network, the posterior density of each element of $\boldsymbol{v}$ is approximated with the Kumaraswamy distribution [31], i.e., $q(v_k|a_k, b_k) = Kumaraswamy(a_k, b_k)$, where $Kumaraswamy(\cdot)$ denotes the Kumaraswamy function with the learnable parameters a_k and b_k. Therefore, the K-L loss $\mathcal{L}_{Beta}$ between the posterior density and the prior density of $\boldsymbol{v}$ is formulated as

$$\begin{aligned}\mathcal{L}_{Beta} &= \sum_{k=1}^{K} D_{KL}(q(v_k|a_k, b_k)||p(v_k)) \\ &= \frac{a_k - \alpha}{a_k}(-\gamma - \Psi(b_k) - \frac{1}{b_k}) + \log a_k b_k + \log B(\alpha, 1) \\ &\quad - \frac{b_k - 1}{b_k},\end{aligned} \tag{6}$$

where γ and $\Psi(\cdot)$ denote the Euler constant and the digamma function, respectively; $B(\cdot)$ is the Beta function [31].

To facilitate the training of LFS, the posterior density of $\mathbf{D}$ is approximated with the Concrete distribution [19, 29]. That is, the posterior density of $\mathbf{D}_i$ can be described as $q(\mathbf{D}_i|\boldsymbol{v}, \mathbf{x}_t^i) = Concrete((\boldsymbol{\pi} + \mathbf{h}^i), \lambda_q)$, where $\boldsymbol{\pi} = [\pi_1, \pi_2, \cdots, \pi_K]$ and $\mathbf{h}^i = [h_1^i, h_2^i, \cdots, h_K^i]$ is the noise; $Concrete(\cdot)$ indicates the Concrete function; λ_q denotes the temperature parameter for the posterior density. Hence, according to [29], the K-L loss $\mathcal{L}_{Bern}$ between the posterior density and the prior density of $\mathbf{D}$ is

$$\begin{aligned}\mathcal{L}_{Bern} &= \sum_{i=1}^{N} D_{KL}(q(\mathbf{D}_i|\boldsymbol{v}, \mathbf{x}_t^i)||p(\mathbf{D}_i|\boldsymbol{v})) \\ &= \sum_{i=1}^{N}(\sum_{k=1}^{K} \log \rho_{\delta_q^{i,k},\lambda_q}(\widetilde{d}_{i,k}) - \sum_{k=1}^{K} \log \kappa_{\delta_p^{i,k},\lambda_p}(\widetilde{d}_{i,k})),\end{aligned} \tag{7}$$

where $\log \rho_{\delta_q^{i,k},\lambda_q}(\widetilde{d}_{i,k})$ and $\log \kappa_{\delta_p^{i,k},\lambda_p}(\widetilde{d}_{i,k})$ denote the log-density functions in the binary concrete case for the posterior density and the prior density, respectively; λ_p is the temperature parameter for the prior density; $\widetilde{d}_{i,k}$ is sampled from $\mathbf{D}$ as $\widetilde{d}_{i,k} = F_\sigma(\frac{1}{\lambda_q}(\log \delta_q^{i,k} + \log U - \log(1 - U)))$, where U obeys a uniform distribution on [0,1]; F_σ denotes the sigmoid function; $\delta_q^{i,k} = \pi_k + h_k^i$ and $\delta_p^{i,k} = \pi_k$ denote the logit probabilities in the posterior density and the prior density, respectively, for the k-th factor in the i-th image.

Similarly, the K-L loss $\mathcal{L}_{Gaus}$ between the posterior density and the prior density of $\mathbf{W}$ is formulated as

$$\begin{aligned}\mathcal{L}_{Gaus} &= \sum_{i=1}^{N} D_{KL}(q(\mathbf{W}_i|\mathbf{x}_t^i)||p(\mathbf{W}_i)) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K}(-\log \sigma_{i,k} + \frac{\sigma_{i,k}^2 + \mu_{i,k}^2}{2} - \frac{1}{2}),\end{aligned} \tag{8}$$

where the posterior density is $q(\mathbf{W}_i|\mathbf{x}_t^i) = \mathcal{N}(\boldsymbol{\mu}_i, diag(\boldsymbol{\sigma}_i^2))$; the mean $\mu_{i,k}$ and the variance $\sigma_{i,k}^2$ respectively denote the k-th elements of $\boldsymbol{\mu}_i \in \mathbb{R}^{1\times K}$ and $\boldsymbol{\sigma}_i^2 \in \mathbb{R}^{1\times K}$, which are learned in LFS.

Therefore, by combining the above three losses in LFS, we have

$$\mathcal{L}_{LFS} = \mathcal{L}_{Beta} + \mathcal{L}_{Bern} + \mathcal{L}_{Gaus}. \tag{9}$$

By obtaining $\mathbf{D}$ and $\mathbf{W}$, $f^u \in \mathbb{R}^{N\times K}$ can be obtained as

$$f^u = \mathbf{D} \odot \mathbf{W}, \tag{10}$$

where $f^u = [f_1^u, f_2^u, \cdots, f_N^u]^{\mathrm{T}}$ represents the extracted label-free disturbance features for all the images in the FER database; $\odot$ denotes the element-wise multiplication.

3.4 Adversarial Training

Inspired by [16], adversarial training is further adopted to enlarge the differences between label-free disturbance features and expression features. We employ a classifier C_{adv}^u (including an FC layer and a parametric ReLU layer) to play an adversarial game with LFS.

Specifically, given an image $\mathbf{x}_t^i$ and its corresponding expression label y_t^i, $f_i^u \in \mathbb{R}^{1\times K}$ denotes the feature extracted from LFS. The adversarial training contains two steps. First, an additional classifier C_{adv}^u is trained to predict expressions given the disturbance feature f_i^u as the input. Note that f_i^u involves some expression information before adversarial training. Hence, it can still be used to predict expressions. Thus, C_{adv}^u is updated by minimizing the classification loss $\mathcal{L}_{cls}^u$ as

$$\mathcal{L}_{cls}^u = -\sum_{i=1}^{N}\sum_{c=1}^{C} \mathbb{1}_{[c=y_t^i]} \log(\mathcal{P}_u(f_i^u)), \tag{11}$$

where $\mathcal{P}_u$ denotes the prediction function of C_{adv}^u.

Second, the feature extractor E_f of LFS is trained to fool C_{adv}^u by maximizing the uncertainty of predictions from C_{adv}^u. Like [16], we define a confusion loss which minimizes the cross-entropy between predictions and a uniform distribution over expression labels as

$$\mathcal{L}_{conf}^u = -\frac{1}{C}\sum_{i=1}^{N} \log(\mathcal{P}_u(f_i^u)). \tag{12}$$

Therefore, the adversarial loss is expressed as

$$\mathcal{L}_{adv}^u = \mathcal{L}_{cls}^u + \mathcal{L}_{conf}^u. \tag{13}$$

Analogously, we also perform adversarial training to pull away label-aware disturbance features from expression features. Similar to Eq. (13), the adversarial loss is formulated as

$$\mathcal{L}_{adv}^l = \mathcal{L}_{cls}^l + \mathcal{L}_{conf}^l, \tag{14}$$

where $\mathcal{L}_{adv}^l$, $\mathcal{L}_{cls}^l$, and $\mathcal{L}_{conf}^l$ respectively denote the adversarial loss, the classification loss, and the confusion loss during adversarial training between a classifier C_{adv}^l and LAS.

3.5 Joint Loss

Based on the above formulations, the joint loss of the proposed D^{3}Net is given as

$$\mathcal{L}_{joint} = \mathcal{L}_{exp} + \eta_1\mathcal{L}_{LAS} + \eta_2\mathcal{L}_{LFS} + \eta_3\mathcal{L}_{adv}^u + \eta_4\mathcal{L}_{adv}^l, \tag{15}$$

where η_1, η_2, η_3, and η_4 represent the balancing weights.

By optimizing the joint loss, D^{3}Net is capable of effectively disentangling the disturbance from facial images and extracting discriminative expression features for FER. It is worth pointing out that we do not impose a constraint to enforce the differences between the features extracted from LFS and LAS since these features are not mutually uncorrelated (e.g., the correlations between some common disturbing factors (such as gender) and some potential disturbing factors (such as hairstyle) can be high).

3.6 Discussions

Conventional FER methods either explicitly alleviate the influence caused by common disturbing factors [26, 32, 36] or implicitly suppress the variations of all disturbing factors [2, 21, 42]. In contrast, D^{3}Net is designed to perform both explicit and implicit disturbance disentanglement by designing LAS and LFS. On the one hand, different from DDL [32] that employs adversarial transfer learning, we adopt the K-L loss to transfer the knowledge from a pretrained model to LAS. Such a way is simple but effective. Moreover, we design LAS to enable the adaptive selection of common disturbing factors according to different characteristics of FER databases. On the other hand, in LFS, we exploit a non-parametric IBP prior to obtain a richer posterior approximation than the commonly used Gaussian posterior approximation. This is more suitable to implicitly disentangle the potential disturbing factors for FER. In addition, most VAE-based UDRL methods [15, 18, 20] simultaneously perform image reconstruction and disentanglement. Although image reconstruction is beneficial to capture the detailed information, it can be detrimental to perform disentanglement. Unlike these methods, we leverage adversarial training to maximize the discrepancy between label-free disturbance features and expression features (instead of using the decoder for image reconstruction), leading to improved classification performance.

4 EXPERIMENTS

4.1 Databases

CK+. The Extended Cohn-Kanade (CK+) database [27] is one of the most popular in-the-lab databases, and it contains 327 video sequences annotated with six basic expressions (i.e., happy, angry, sad, surprise, fear, and disgust) and contempt. Since CK+ does not offer the training, validation, and test sets, the last three peak frames from each sequence are selected to construct the dataset. The whole dataset is then split into 10 subsets based on the identity, where the subjects are mutually exclusive in any two subsets. Following [32, 42], we adopt the popular 10-fold subject-independent cross-validation in this paper.

Oulu-CASIA. The Oulu-CASIA database [48] contains 2,880 image sequences labeled with six basic expressions. The images are captured with two imaging systems (i.e., near-infrared (NIR) and visible light (VIS)), under three different illumination conditions. Following the settings in [42], we choose the last three peak frames in each sequence captured by the VIS system under the strong illumination condition for experiments. Similar to the CK+ database, we perform the 10-fold subject-independent cross-validation.

MMI. The MMI database [35] is a laboratory-controlled database and involves challenging inter-personal variations. It includes 205 image sequences labeled with six basic facial expressions. Similar to the CK+ database, we select the three peak frames in each frontal sequence for 10-fold subject-independent cross-validation.

Table 1: The details of the baseline and 5 variants of D^{3}Net.

Methods	B_e	LAS	C^l_{adv}	LFS	C^u_{adv}
baseline	√				
D^{3}Net-LA	√	√	√		
D^{3}Net-LF	√			√	√
D^{3}Net-LA_AT	√	√			
D^{3}Net-LF_AT	√			√	
D^{3}Net	√	√	√	√	√

B_e denotes the expression branch.

RAF-DB. The Real-world Affective Face Database (RAF-DB) [25] is an in-the-wild database that contains 15,339 images labeled with six basic facial expressions and one neutral expression. Besides, it also provides labels of gender, race, and age. RAF-DB is divided into 12,271 training samples and 3,068 test samples.

SFEW. The Static Facial Expressions in the Wild (SFEW) [7] is created by selecting the static frames from the AFEW database. SFEW 2.0 [8] is the most commonly used version, where contains 958 images for training and 436 images for validation. Each image is labeled with one of six basic expressions or the neutral expression.

4.2 Implementation Details

In our experiments, all the images are aligned and cropped to extract the facial regions. Then, they are resized to 256 × 256 pixels and cropped to 224 × 224 pixels. Furthermore, we apply the horizontal flip to the cropped images. Similar to [37, 38], the shared backbone network is based on ResNet-18 [17], where we remove the last FC layer and change the output size of the last pooling layer. In this paper, ResNet-18 is pretrained on the MS-Celeb-1M face database [14]. As a result, the shared backbone network outputs a 2,048-dimensional feature vector. The first FC layers in LAS, LFS, and expression branch have the size of 512. The dimensions of each f^a_i and f^u_i are set as 150 (i.e., $D = K = 150$). The weights in Eq. (15) are empirically set as $\eta_1 = 0.1$, $\eta_2 = 0.1$, $\eta_3 = 1$, and $\eta_4 = 1$. In LFS, we initialize the parameter α of Beta distribution with 20. In LAS, same as DDL [32], the labels of illumination, pose, and identity in the Multi-PIE database and those of gender, race, and age in the RAF-DB database are employed to obtain a pretrained model. Therefore, the value of M in Eq. (3) is 6. The value of M' is adaptively set (4 for Oulu-CASIA, 5 for both CK+ and MMI, and 6 for both SFEW and RAF-DB) according to the experimental results of DDL.

All experiments are implemented by Pytorch and run on NVIDIA GTX TITAN XP. We train the networks for 40 epochs using the Adam optimizer [22] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The initial learning rate is 0.001 and decays by 0.1 after 10, 18, 25, and 32 epochs. The batch size is set as 16 for all the five databases.

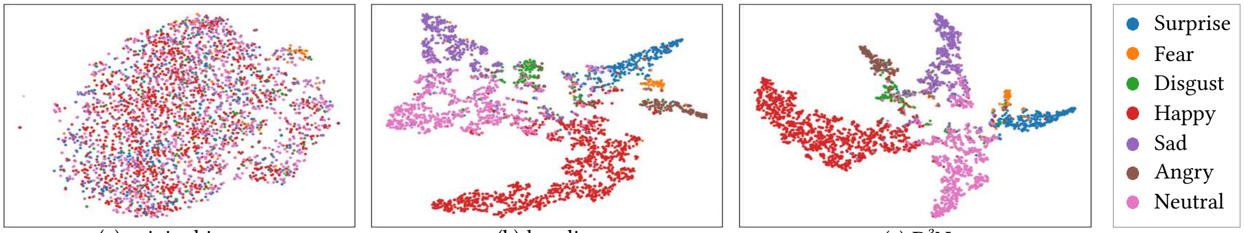
4.3 Ablation Studies

In this subsection, we perform ablation studies to show the influence of the key components of D^{3}Net (including LAS, LFS, and adversarial training) on the performance. We evaluate one baseline method and 5 variants of D^{3}Net. The details of these methods are summarized in Table 1. The comparison results are given in Table 2.

Influence of Label-Aware Sub-Branch (LAS). As shown in Table 2, D^{3}Net-LA achieves better performance than the baseline method (about 1.35%, 5.91%, 2.92%, 1.41%, and 3.44% improvements

Table 2: The classification accuracy (%) obtained by different methods in ablation studies. The best results are boldfaced.

Methods	CK+	MMI	Oulu-CASIA	RAF-DB	SFEW
baseline	97.57	79.23	84.72	86.86	56.65
D^{3}Net-LA	98.92	85.14	87.64	88.27	60.09
D^{3}Net-LF	98.72	85.12	87.29	88.23	60.32
D^{3}Net-LA_AT	98.61	83.54	86.67	87.81	58.49
D^{3}Net-LF_AT	98.54	83.40	86.81	87.61	58.72
D^{3}Net	**99.52**	**86.30**	**89.24**	**88.79**	**62.16**



(a) original images

(b) baseline

(c) D^{3}Net

Figure 3: Feature visualization on (a) the original images, and the models trained by (b) baseline and (c) D^{3}Net on RAF-DB.

in terms of recognition accuracy on CK+, MMI, Oulu-CASIA, RAF-DB, and SFEW, respectively). Similarly, D^{3}Net obtains higher recognition accuracy than D^{3}Net-LF. This demonstrates that LAS, which exploits the available label information from the face databases to explicitly disentangle common disturbing factors in the FER databases, is helpful to perform FER.

Note that, according to the observations in DDL, we select different common disturbing factors to extract disturbance-aware features for different FER databases. More specifically, we choose six common disturbing factors (gender, age, race, identity, illumination, and pose) for two in-the-wild databases. We use five disturbing factors (gender, age, race, identity, and illumination) and four disturbing factors (gender, age, race, and identity) for CK+ and MMI, and Oulu-CASIA, respectively. By transfer learning, we apply the knowledge (referring to common disturbing factors) learned from the face databases to the target FER database, enabling the effective removal of the disturbance caused by these disturbing factors. This successfully alleviates the training difficulty from the lack of label information for common disturbing factors in most FER databases.

Influence of Label-Free Sub-Branch (LFS). From Table 2, compared with the baseline method, D^{3}Net-LF gives better recognition accuracy on both in-the-lab and in-the-wild databases. To be specific, for three in-the-lab databases, D^{3}Net-LF improves the accuracy of the baseline method by 1.15%, 5.89%, and 2.57% on CK+, MMI, and Oulu-CASIA, respectively. For two in-the-wild databases, D^{3}Net-LF respectively obtains 1.37% and 3.67% improvements on RAF-DB and SFEW. Moreover, D^{3}Net outperforms D^{3}Net-LA on five databases. These results show the effectiveness of LFS.

On the one hand, we introduce the IBP prior to model the distribution of potential disturbing factors. Such a manner is advantageous to encode label-free disturbance features in an unsupervised way. On the other hand, we capitalize on adversarial training to enforce label-free disturbance features to be dissimilar from expression features. This is beneficial to extract effective disturbance features, thus capturing the information of potential disturbing factors.

Influence of Adversarial Training. In order to further show the importance of adversarial training, we also evaluate the performance of the models trained without using the adversarial training loss $\mathcal{L}^{l}_{adv}$ and $\mathcal{L}^{u}_{adv}$, which are denoted as D^{3}Net-LA_AT and D^{3}Net-LF_AT, respectively. The results are shown in Table 2.

D^{3}Net-LA_AT achieves worse performance than D^{3}Net-LA in all the databases. This indicates that employing adversarial training to label-aware disturbance features is very helpful to perform disentanglement of disturbance in LAS. In D^{3}Net-LA_AT, label-aware disturbance features are learned by leveraging transfer learning on a pretrained model. Without adversarial training, label-aware disturbance features cannot be completely distinguished from expression features, thereby leading to inferior disentanglement ability.

Compared with D^{3}Net-LF, D^{3}Net-LF_AT also obtains worse recognition accuracy on all the databases. By employing adversarial training in D^{3}Net-LF, the distances between label-free disturbance features and expression features are enforced to be far away from each other. Therefore, we are able to capture the label-free disturbance information, thus improving the performance.

Feature visualization. To further verify that employing both explicit and implicit disturbance disentanglement plays a vital role in extracting discriminative expression features, we use t-SNE [28] to visualize expression features on RAF-DB, as shown in Figure 3.

From the feature distribution of the original images, we are difficult to distinguish different expressions. For the baseline method, the differences between the features from different expressions are not distinct. In Figure 3(b), features from disgust, sad, angry, and neutral are mixed together, which easily leads to classification errors. This can be ascribed to the fact that the features extracted by the baseline method are easily affected by the disturbance due to various disturbing factors. In other words, the expression information is seriously entangled with the disturbance information for the features obtained by the baseline method. In contrast, compared

Table 3: Performance comparisons on the in-the-lab databases (i.e., CK+, MMI, and Oulu-CASIA) in terms of recognition accuracy (%). The best results are boldfaced.

Method	CK+	MMI	Oulu-CASIA
IACNN [30]	95.37	71.55	–
PHRNN-MSCNN [47]	98.50	81.18	86.25
FN2EN [9]	98.60*	–	87.71
DLP-CNN [24]	95.78*	78.46	–
IPA2LT [44]	92.45*	65.61	61.49
DeRL [42]	97.30	73.23	88.00
ADFL [2]	98.17	77.51	87.50
TDGAN [41]	97.53±2.03*	–	–
DDL [32]	99.16	83.67	88.26
D^{3}Net (proposed)	**99.52**	**86.30**	**89.24**

* denotes that six basic expressions are classified in CK+.

Table 4: Performance comparisons on the in-the-wild databases (i.e., RAF-DB and SFEW) in terms of recognition accuracy (%). The best results are boldfaced.

Method	RAF-DB	SFEW
DLP-CNN [25]	84.13	51.05
IACNN [30]	–	50.98
SPDNet [1]	87.00	58.14
IPA2LT [44]	86.77	58.29
IPFR [36]	–	57.10†
TDGAN [41]	81.91±1.18	–
RAN [38]	86.90	56.40‡
SCN [37]	87.03	–
DDL [32]	87.71	59.86
D^{3}Net (proposed)	**88.79**	**62.16**

† indicates extra data are used during training;

‡ represents a naive model fusion by averaging the scores of ResNet18 and VGG16 [38].

with the baseline method, D^{3}Net shows better inter-class separability and intra-class compactness (see Figure 3(c)). Moreover, the features corresponding to some similar expressions (i.e., sad, neutral, and angry) are also more distinguishable. This is because D^{3}Net is capable of fully suppressing different types of disturbing factors involved in facial images by the disturbance branch. In a word, the proposed D^{3}Net can extract discriminative expression features by performing explicit and implicit disturbance disentanglement.

4.4 Comparison with State-of-the-Art FER Methods

Table 3 and Table 4 report the results obtained by all the competing methods on in-the-lab databases (i.e., CK+, MMI, and Oulu-CASIA) and in-the-wild databases (i.e., RAF-DB and SFEW), respectively.

In Table 3, we can observe that D^{3}Net and DDL outperform the other competing FER methods on all the in-the-lab databases. This is because multiple common disturbing factors are simultaneously suppressed by these two methods. In contrast, IACNN only considers the influence caused by identity. Compared with DDL that only performs explicit disentanglement of limited common disturbing factors, D^{3}Net performs both explicit and implicit disturbance disentanglement by using LAS and LFS. Hence, D^{3}Net achieves higher recognition accuracy than DDL. This demonstrates the significance of taking into account potential disturbing factors in FER. TDGAN, PHRNN-MSCNN, and DeRL show good performance on CK+, MMI, and Oulu-CASIA. However, TDGAN and DeRL extract discriminative expression features by implicitly separating the disturbance information. Note that only six basic expressions in CK+ are classified by TDGAN, while seven expressions are predicted by our D^{3}Net. PHRNN-MSCNN focuses on the appearance variations and identity differences from sequence data, ignoring many other potential disturbing factors.

As shown in Table 4, we compare the proposed method with nine state-of-the-art FER methods on two in-the-wild databases. Clearly, D^{3}Net obtains better performance than the other competing FER methods on both RAF-DB and SFEW. Although DLP-CNN, SPDNet, IPA2LT, and SCN achieve excellent FER performance, these methods do not explicitly consider the disturbance in facial images. On the contrary, some common disturbing factors (such as illumination and pose) are taken into account in D^{3}Net. By performing transfer learning in LAS, D^{3}Net is capable of explicitly alleviating the negative influence of these disturbing factors. Therefore, the above results show the importance of disturbance disentanglement for performing effective expression recognition.

Among the disturbance-disentangled based FER methods, IACNN, IPFR, RAN, and DDL only consider limited common disturbing factors. However, these methods ignore the influence of potential disturbing factors, thus resulting in inferior recognition accuracy. TDGAN is proposed to implicitly disentangle variations by GAN, but it may not well cope with common disturbing factors. In contrast, D^{3}Net effectively disentangles the disturbance of various disturbing factors, achieving the best performance among all the competing methods.

5 CONCLUSION

In this paper, we develop a novel D^{3}Net, mainly consisting of an expression branch and a disturbance branch, to address the disturbance in facial images for effective FER. In the disturbance branch, two sub-branches (i.e., LAS and LFS) are respectively designed to perform explicit and implicit disturbance disentanglement. In LAS, we leverage transfer learning to capture the information of common disturbing factors. Meanwhile, in LFS, we introduce the IBP prior to model the distribution of potential disturbing factors. Moreover, adversarial training is employed to effectively enlarge the differences between disturbance features and expression features. By jointly training the disturbance branch and the expression branch on the shared backbone network, we are able to extract discriminative expression features for FER. A large number of experiments on both in-the-lab and in-the-wild databases have shown the superiority of D^{3}Net in comparison with several state-of-the-art FER methods.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grants 62071404 and 61872307, by the Natural Science Foundation of Fujian Province under Grant 2020J01001, and by the Youth Innovation Foundation of Xiamen City under Grant 3502Z20206046.

REFERENCES

[1] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. 2018. Covariance pooling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 367–374.

[2] Mengchao Bai, Weicheng Xie, and Linlin Shen. 2019. Disentangled feature based adversarial learning for facial expression recognition. In *Proceedings of the IEEE International Conference on Image Processing*. 31–35.

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828.

[4] Jiawei Chen, Janusz Konrad, and Prakash Ishwar. 2018. Vgan-based image representation learning for privacy-preserving facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1570–1579.

[5] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. 2018. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942* (2018).

[6] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. 2016. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (2016), 1548–1568.

[7] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2011. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2106–2112.

[8] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. 2015. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the ACM on International Conference on Multimodal Interaction*. 423–426.

[9] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. 2017. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. 118–126.

[10] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *arXiv preprint arXiv:1406.2661* (2014).

[11] Thomas L Griffiths and Zoubin Ghahramani. 2005. Infinite latent feature models and the Indian buffet process. In *Proceedings of Advances in Neural Information Processing Systems*, Vol. 18. 475–482.

[12] Thomas L Griffiths and Zoubin Ghahramani. 2011. The Indian buffet process: An introduction and review. *Journal of Machine Learning Research* 12, 4 (2011), 1185–1224.

[13] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. 2010. Multi-PIE. *Image and Vision Computing* 28, 5 (2010), 807–813.

[14] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision*. 87–102.

[15] Prashnna Gyawali, Zhiyuan Li, Cameron Knight, Sandesh Ghimire, B Milan Horacek, John Sapp, and Linwei Wang. 2019. Improving disentangled representation learning with the beta bernoulli process. In *Proceedings of the IEEE International Conference on Data Mining*. 1078–1083.

[16] Marah Halawa, Manuel Wöllhaf, Eduardo Vellasques, Urko SánchezSanz, and Olaf Hellwich. 2020. Learning disentangled expression representations from facial images. *arXiv preprint arXiv:2008.07001* (2020).

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[18] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations*.

[19] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[20] Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *Proceedings of the International Conference on Machine Learning*. 2649–2658.

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

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

[23] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[24] Shan Li and Weihong Deng. 2018. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing* 28, 1 (2018), 356–370.

[25] Shan Li, Weihong Deng, and JunPing Du. 2017. 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*. 2852–2861.

[26] Yuanyuan Liu, Jiabei Zeng, Shiguang Shan, and Zhuo Zheng. 2018. Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. 458–465.

[27] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 94–101.

[28] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008), 2579–2605.

[29] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712* (2016).

[30] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. 2017. Identity-aware convolutional neural network for facial expression recognition. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. 558–565.

[31] Eric Nalisnick and Padhraic Smyth. 2016. Stick-breaking variational autoencoders. *arXiv preprint arXiv:1605.06197* (2016).

[32] Delian Ruan, Yan Yan, Si Chen, Jing-Hao Xue, and Hanzi Wang. 2020. Deep disturbance-disentangled learning for facial expression recognition. In *Proceedings of the ACM International Conference on Multimedia*. 2833–2841.

[33] Rachit Singh, Jeffrey Ling, and Finale Doshi-Velez. 2017. Structured variational autoencoders for the beta-bernoulli process. In *Proceedings of NIPS Workshop on Advances in Approximate Bayesian Inference*.

[34] Yee Whye Teh, Dilan Grür, and Zoubin Ghahramani. 2007. Stick-breaking construction for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. 556–563.

[35] Michel Valstar and Maja Pantic. 2010. Induced disgust, happiness and surprise: An addition to the MMI facial expression database. In *Proceedings of the International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*. 65–70.

[36] Can Wang, Shangfei Wang, and Guang Liang. 2019. Identity-and pose-robust facial expression recognition through adversarial feature learning. In *Proceedings of the ACM International Conference on Multimedia*. 238–246.

[37] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. 2020. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6896–6905.

[38] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. 2020. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing* 29, 1 (2020), 4057–4069.

[39] Wenxuan Wang, Qiang Sun, Yanwei Fu, Tao Chen, Chenjie Cao, Ziqi Zheng, Guoqiang Xu, Han Qiu, Yu-Gang Jiang, and Xiangyang Xue. 2019. Comp-GAN: Compositional generative adversarial network in synthesizing and recognizing facial expression. In *Proceedings of the ACM International Conference on Multimedia*. 211–219.

[40] Bin Xia and Shangfei Wang. 2020. Occluded facial expression recognition with step-wise assistance from unpaired non-occluded images. In *Proceedings of the ACM International Conference on Multimedia*. 2927–2935.

[41] Siyue Xie, Haifeng Hu, and Yizhen Chen. 2020. Facial expression recognition with two-branch disentangled generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology* (2020), 1–1.

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

[43] Huiyuan Yang, Zheng Zhang, and Lijun Yin. 2018. Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. 294–301.

[44] Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2018. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European Conference on Computer Vision*. 222–237.

[45] Feifei Zhang, Tianzhu Zhang, Qirong Mao, Lingyu Duan, and Changsheng Xu. 2018. Facial expression recognition in the wild: A cycle-consistent adversarial attention transfer approach. In *Proceedings of the ACM International Conference on Multimedia*. 126–135.

[46] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. 2020. Geometry guided pose-invariant facial expression recognition. *IEEE Transactions on Image Processing* 29 (2020), 4445–4460.

[47] Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. 2017. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing* 26, 9 (2017), 4193–4203.

[48] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. 2011. Facial expression recognition from near-infrared videos. *Image and Vision Computing* 29, 9 (2011), 607–619.