Deep Disturbance-Disentangled Learning for Facial Expression Recognition

Delian Ruan
Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen, China
delianruan@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

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

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

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

To achieve effective facial expression recognition (FER), it is of great importance to address various disturbing factors, including pose, illumination, identity, and so on. However, a number of FER databases merely provide the labels of facial expression, identity, and pose, but lack the label information for other disturbing factors. As a result, many methods are only able to cope with one or two disturbing factors, ignoring the heavy entanglement between facial expression and multiple disturbing factors. In this paper, we propose a novel Deep Disturbance-disentangled Learning (DDL) method for FER. DDL is capable of simultaneously and explicitly disentangling multiple disturbing factors by taking advantage of multi-task learning and adversarial transfer learning. The training of DDL involves two stages. First, a Disturbance Feature Extraction Model (DFEM) is pre-trained to perform multi-task learning for classifying multiple disturbing factors on the large-scale face database (which has the label information for various disturbing factors). Second, a Disturbance-Disentangled Model (DDM), which contains a global shared sub-network and two task-specific (i.e., expression and disturbance) sub-networks, is learned to encode the disturbance-disentangled information for expression recognition. The expression sub-network adopts a multi-level attention mechanism to extract expression-specific features, while the disturbance sub-network leverages adversarial transfer learning to extract disturbance-specific features based on the pre-trained DFEM. Experimental results on both the in-the-lab FER databases (including CK+, MMI, and Oulu-CASIA) and the in-the-wild FER databases (including RAF-DB and SFEW) demonstrate the superiority of our proposed method compared with several state-of-the-art methods.

(a) inter-class similarity (b) intra-class difference

Figure 1: Facial expression images with (a) high inter-class similarity (the images are from the CK+ database [20]) and (b) high intra-class difference (the images are from the SFEW database [5]).

CCS CONCEPTS

• **Computing methodologies** → **Image representations**.

KEYWORDS

Facial expression recognition, Multi-task learning, Adversarial transfer learning, Multi-level attention.

ACM Reference Format:
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 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413907

1 INTRODUCTION

Facial expression is an effective communicative signal for human beings to express their inner states [3]. Facial expression recognition (FER), as an important and fundamental task in computer vision

*Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413907

and multimedia, has attracted much attention due to its variety of applications in security, digital entertainment, driver monitoring, and so on [38, 39].

Over the past few years, inspired by the outstanding performance of deep learning [15], convolutional neural network (CNN) based FER methods [17, 22, 23, 25, 29, 32, 35] have shown promising recognition accuracy. Despite significant progress, FER is still a very challenging task. In particular, facial expression images are often intertwined with various disturbing factors, such as pose, identity, illumination, age, gender, etc. These disturbing factors have a substantial influence on the natural appearance of facial images. As shown in Figure 1, facial expression images show significant inter-class similarities and intra-class differences because of different disturbing factors. For each row in Figure 1(a), the two images of different expressions exhibit high similarity due to the same illumination and identity. For the images in Figure 1(b), the four images of the same expression show great differences due to variations in gender, age, race, identity, illumination, and pose. Obviously, these disturbing factors seriously interfere with the extraction of expression-related information.

Hence, it is critical to disentangle the disturbance information while retaining the expression-related information for deep features. Many CNN based FER methods [23, 36] have been developed to implicitly suppress the disturbance information in facial expression images. Generally, the training of CNN requires a large amount of labeled data to ensure excellent performance. However, many FER databases only provide limited training data. Therefore, the CNN models obtained by these methods may not effectively alleviate the influence of various disturbing factors, given limited training data.

Recently, some disturbance-disentangled based FER methods [22, 29, 39] have been proposed to explicitly disentangle the disturbing factors for FER. Nevertheless, many FER databases merely provide the labels of facial expression and identity (or pose), and lack the label information for other disturbing factors. As a result, these methods usually consider only a few disturbing factors, since manually labeling various disturbing factors is time-consuming. The performance of these methods is still far from being satisfactory.

Fortunately, there exist some large-scale face databases containing a large number of facial images with the label information for different disturbing factors (e.g., Multi-PIE [8] offers the labels of identity, pose, and illumination; RAF-DB [17] gives the labels of gender, race, and age). In this paper, we exploit the available disturbance label information (i.e., the labels of disturbing factors) in these large-scale face databases to perform adversarial transfer learning for identifying expressions on the disturbance unlabeled FER databases. As a consequence, the problems of limited training data and the lack of disturbance labels can be effectively addressed.

To be specific, we propose an effective FER method termed Deep Disturbance-disentangled Learning (DDL), which is capable of disentangling multiple disturbing factors from facial expression images and learning expression-specific features, by taking advantage of multi-task learning and adversarial transfer learning. The training of DDL involves a two-stage learning procedure. First, a Disturbance Feature Extraction Model (DFEM) is pre-trained to classify multiple disturbing factors. Second, a Disturbance-Disentangled Model (DDM) is trained to disentangle the disturbance information and obtain expression-specific features. The DDM is comprised of a global shared sub-network and two task-specific (i.e., expression and disturbance) sub-networks. The expression sub-network is designed to extract expression-specific features, while the disturbance sub-network aims to extract disturbance-specific features based on the pre-trained DFEM. In particular, a multi-level attention mechanism is adopted to exploit both low-level spatial features and high-level semantic features in the expression sub-network.

In summary, the main contributions of our work include:

(1) We propose a novel DDL method, which consists of two models (i.e., a DFEM and a DDM), for effective FER. The proposed method is able to simultaneously disentangle multiple disturbing factors and encode the expression-related information for expression recognition. To the best of our knowledge, the proposed method is the first work to harness the available disturbance label information from the large-scale face database to perform transfer learning on the disturbance unlabeled FER database.

(2) We elaborately design two task-specific sub-networks in the DDM. For the expression sub-network, a multi-level attention mechanism is developed to extract expression-specific features. For the disturbance sub-network, adversarial transfer learning is adopted to learn disturbance-specific features. We jointly train the two sub-networks based on the global shared sub-network from low-level layers to high-level layers.

(3) The proposed DDL is extensively evaluated on both the in-the-lab and in-the-wild FER databases. Experimental results show that our proposed method consistently outperforms several state-of-the-art methods, which can verify the importance of disturbance disentangling for effective FER.

2 RELATED WORK

In this section, we respectively discuss the existing works about CNN based FER methods, disturbance-disentangled based FER methods, and attention mechanisms, which are closely related to our proposed method.

2.1 CNN Based FER Methods

Due to its powerful representation capability, CNN has attracted significant attention in the areas of multimedia and computer vision. Currently, the CNN based FER methods [15] have achieved state-of-the-art performance. For example, Yu and Zhang [36] propose an ensemble of CNNs, which shows promising results in the EmotiW challenge. Mollahosseini *et al.* [23] develop a network consisting of two convolutional layers and four Inception layers [27] for FER. Hu *et al.* [12] propose a supervised scoring ensemble (SSE) method based on ResNet [9], where the supervision signal is not only used for the deep layers, but also used for the intermediate and shallow layers.

These CNN based FER methods implicitly alleviate the influence of various disturbing factors involved in facial expression images. Note that CNN usually requires a large number of training data to learn powerful feature representations. However, many FER databases do not have sufficient training samples. Therefore, one potential problem of these methods is that the trained CNN models are not robust to handle various disturbing factors.

2.2 Disturbance-Disentangled Based FER Methods

Recently, some methods have been proposed to explicitly perform disturbance disentangling for FER. For example, Zhang *et al.* [39] propose a generative adversarial network (GAN) based pose-invariant method for simultaneous facial image synthesis and FER, by exploiting the relationship between different poses and expressions. Therefore, the influence of pose variations on FER is effectively mitigated. Meng *et al.* [22] propose an identity-aware convolutional neural network (IACNN) method to alleviate the variations caused by the facial identity, where an identity-sensitive contrastive loss is adopted to learn the identity-related information. Wang *et al.* [29] propose an adversarial feature learning method to disentangle the disturbances caused by pose and identity.

The above disturbance-disentangled based FER methods require the labels of disturbing factors in the FER databases. Unfortunately, many FER databases only provide the labels of facial expression and some facial attributes (such as identity and pose), but the label information for other disturbing factors is not available. Thus, these methods are only able to handle one or two disturbing factors. Different from the above methods, our method can effectively disentangle multiple disturbing factors from facial expression images by capitalizing on the disturbance label information available in large-scale face databases to perform transfer learning.

2.3 Attention Mechanisms

In recent years, some attention mechanism based CNN methods have been developed in a variety of tasks, such as fine-grained recognition [7, 11, 43], image caption [34], person re-identification [31], and human pose estimation [2].

On the one hand, psychologists have shown that salient facial regions (such as mouth, nose, and eyes) play a crucial role for FER [26]. On the other hand, attention mechanisms have shown a great capability to select salient features. Therefore, attention mechanisms are beneficial to predict facial expressions. For instance, Xie *et al.* [32] propose a deep attentive multi-path CNN (DAM-CNN) method, where a spatial attention mechanism is adopted to obtain salient regions. Wang *et al.* [30] propose a novel region attention network (RAN) to adaptively capture salient facial regions for occlusion-invariant and pose-invariant FER. In general, these attention mechanism based FER methods leverage high-level semantic features of CNN for expression recognition.

As a matter of fact, both high-level features and low-level features of CNN are advantageous to improve the FER performance. Therefore, in this paper, unlike previous methods, we adopt a multi-level attention mechanism, which aggregates the attentive features from different layers of the network. The attention mechanism effectively exploits both the spatial-aware and semantic-aware information to extract discriminative features for identifying facial expressions.

3 PROPOSED METHOD

In this section, we introduce the proposed DDL method in detail. An overview of the proposed method is first introduced. Each component of the proposed method is then described in detail. Finally, some discussions about the proposed method are given.

3.1 Overview

An overview of the proposed DDL method is shown in Figure 2. The training of DDL involves two stages. In the first stage, a Disturbance Feature Extraction Model (DFEM) is pre-trained to simultaneously classify various disturbing factors (such as gender, race, and age) using the disturbance labeled face database. In this way, the features extracted by the DFEM effectively encode the disturbance information about these disturbing factors. In the second stage, a Disturbance-Disentangled Model (DDM) is learned to perform FER on the disturbance unlabeled FER database. The DDM consists of a global shared sub-network, an expression sub-network, and a disturbance sub-network. The expression sub-network adopts a multi-level attention mechanism to extract expression-specific features. The disturbance sub-network extracts disturbance-specific features by exploiting adversarial transfer learning. During training, by optimizing different loss functions for the expression and disturbance sub-networks, expression-specific features and disturbance-specific features are respectively extracted based on the common global shared features, and thus the disturbance can be explicitly disentangled.

3.2 Disturbance Feature Extraction Model

The network architecture of the DFEM consists of the shared layers and the task-specific layers for classifying different disturbing factors, as shown in Figure 2(a).

Specifically, a facial image is first fed to several shared layers to obtain high-level shared features. In this paper, we adopt the PreAct ResNet-18 [10] as the shared layers. Each task-specific layer consists of a fully-connected layer to extract discriminative features for classifying a disturbing factor. In this way, the features obtained by the shared layers are ensured to encode the disturbance information.

Given a disturbance labeled face database, we have a training set $\mathbf{T}^l$ with R images and their corresponding labels: $\mathbf{T}^l = \{\mathbf{x}_i^l, \mathbf{y}_i\}_{i=1}^R$, where $\mathbf{x}_i^l$ denotes the i-th training image and $\mathbf{y}_i$ is an M-dimensional vector consisting of the labels of the disturbing factors. M denotes the number of disturbing factors. The optimization problem of the DFEM is expressed as:

$$\underset{\mathbf{w}_c, \{\mathbf{w}_j\}_{j=1}^M}{\arg\min} \sum_{i=1}^{R} \sum_{j=1}^{M} \mathcal{L}_{CE}^j(y_i^j, \mathcal{F}_j(\mathbf{x}_i^l, \mathbf{w}_c, \mathbf{w}_j)), \tag{1}$$

where the network parameter $\mathbf{w}_c$ controls feature sharing among all the disturbing factors, and the network parameter $\mathbf{w}_j$ controls the update of the features with respect to each disturbing factor; $\mathcal{F}_j(\cdot,\cdot,\cdot)$ represents the prediction function for the j-th disturbing factor, given the input $\mathbf{x}_i$, the network parameters $\mathbf{w}_c$ and $\mathbf{w}_j$; y_i^j denotes the label of the i-th image corresponding to the j-th disturbing factor; $\mathcal{L}_{CE}^j(\cdot,\cdot)$ represents the cross-entropy (CE) loss between the result estimated by $\mathcal{F}_j$ and the corresponding ground truth label y_i^j. The formula of the CE loss is defined as:

$$\mathcal{L}_{CE}^j = -\sum_{k=1}^{K^j} \Pi_{[k=y_i^j]} \log(\mathcal{F}_j(\mathbf{x}_i^l, \mathbf{w}_c, \mathbf{w}_j)), \tag{2}$$

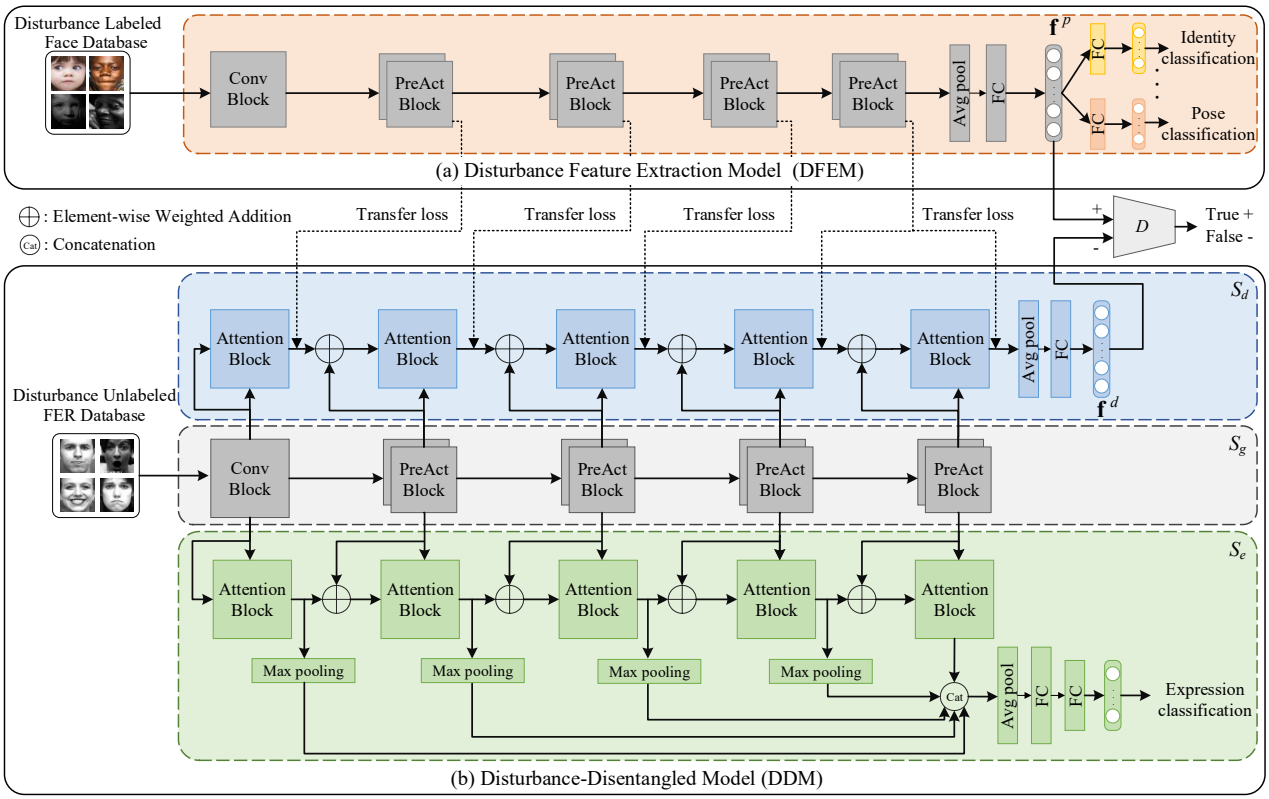

Figure 2: An overview of the proposed DDL method. DDL involves two stages. (a) Pre-training a DFEM consisting of the shared layers and the task-specific layers. The DFEM extracts features for identifying multiple disturbing factors. (b) Training a DDM consisting of a global shared sub-network (S_g), an expression sub-network (S_e), and a disturbance sub-network (S_d). The DDM extracts expression-specific features by disentangling the disturbance information.

where $\log(\cdot)$ denotes the logarithm function; K^j denotes the class number of the j-th disturbing factor; $\Pi_{[k=y_i^j]}$ outputs 1 when $k = y_i^j$, and 0 otherwise.

3.3 Disturbance-Disentangled Model

Based on the pre-trained DFEM on the large-scale face database, a DDM is trained to explicitly disentangle the disturbances from facial images on the disturbance unlabeled FER database. The network architecture of the DDM consists of a global shared sub-network and two task-specific sub-networks, as given in Figure 2(b).

Global Shared Sub-network. The global shared sub-network (denoted as S_g) is designed based on the PreAct ResNet-18 [10], where we remove the final average pooling layer and the fully-connected layer. S_g extracts the global shared features of the input image.

Task-specific Sub-networks. The expression sub-network (denoted as S_e) consists of a set of attention blocks, followed by an average pooling layer and two fully-connected layers. S_e is designed to learn expression-specific features by applying the attention blocks to the global shared sub-network S_g. Here, the attention block generates a soft attention mask, which can indicate the importance of each position in the feature map. Moreover, a multi-level attention mechanism is employed to exploit features at different levels of the network.

Given a disturbance unlabeled FER database, we have a training set $\mathbf{T}^u$ with N images and their corresponding labels: $\mathbf{T}^u = \{\mathbf{x}_i^u, y_i\}_{i=1}^N$, where $\mathbf{x}_i^u$ denotes the i-th training image and y_i is the expression label corresponding to $\mathbf{x}_i^u$. The goal of S_e is to optimize the following problem:

$$\arg\min_{\mathbf{w}_g, \mathbf{w}_e} \sum_{i=1}^{N} \mathcal{L}_{CE}(y_i, \mathcal{F}_e(\mathbf{x}_i^u, \mathbf{w}_g, \mathbf{w}_e)), \tag{3}$$

where $\mathbf{w}_g$ denotes the network parameter in the global shared sub-network S_g; $\mathbf{w}_e$ denotes the network parameter in the expression sub-network S_e; $\mathcal{F}_e(\cdot,\cdot,\cdot)$ denotes the prediction function; and $\mathcal{L}_{CE}$ indicates the CE loss between the ground truth expression label y_i and the predicted result by $\mathcal{F}_e$, which is expressed as:

$$\mathcal{L}_{CE} = -\sum_{k=1}^{K} \Pi_{[k=y_i]} \log(\mathcal{F}_e(\mathbf{x}_i^u, \mathbf{w}_g, \mathbf{w}_e)), \tag{4}$$

where K is the number of expression categories.

Analogously, the disturbance sub-network (denoted as S_d) also contains a set of attention blocks, followed by an average pooling layer and a fully-connected layer. The goal of S_d is to learn the

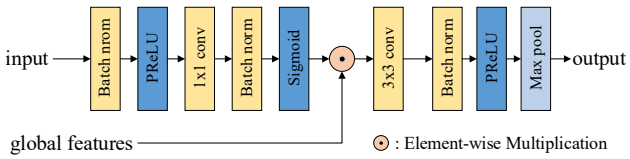

Figure 3: The network architecture of the attention block.

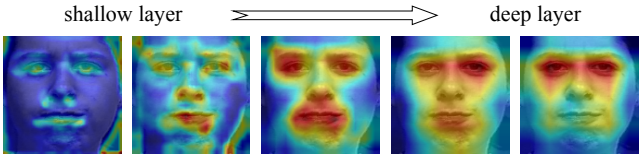

Figure 4: Visualization of the output feature maps from different levels of the S_e sub-network.

disturbance-specific features ($\mathbf{f}^d$), whose distribution is as similar as possible to that of the features ($\mathbf{f}^p$) extracted by the pre-trained DFEM. To achieve this, we take advantage of adversarial transfer learning. In particular, S_d and the discriminator D play an adversarial game, where S_d tries to minimize the divergence of the feature distributions between $\mathbf{f}^d$ and $\mathbf{f}^p$, while D aims to distinguish $\mathbf{f}^d$ from $\mathbf{f}^p$. The objective of adversarial training is formulated as:

$$\min_{D} \max_{S_d} \mathcal{L}_{AD}(S_d, D), \tag{5}$$

where the adversarial loss $\mathcal{L}_{AD}$ is defined as follows:

$$\mathcal{L}_{AD} = -\mathbb{E}[\log(D(\mathbf{f}^p))] - \mathbb{E}[\log(1 - D(\mathbf{f}^d))]. \tag{6}$$

To facilitate the transfer of prior knowledge from the pre-trained DFEM to S_d, it is natural that the distributions of both the final output features and the intermediate attention maps of S_d are statistically close to those of the pre-trained DFEM. Therefore, we also apply the attention transfer [38], which has been proven to be effective in bridging the gap between the source domain and the target domain, by transferring attention knowledge. The attention transfer loss is expressed as:

$$\mathcal{L}_{AT} = \sum_{j=1}^{L} || \frac{\mathbf{q}_d^j}{||\mathbf{q}_d^j||_2} - \frac{\mathbf{q}_p^j}{||\mathbf{q}_p^j||_2} ||_2, \tag{7}$$

where $\mathbf{q}_d^j$ and $\mathbf{q}_p^j$ are the j-th attention maps pair associated with the S_d and the pre-trained DFEM in the vectorized forms; L denotes the number of attention blocks. In this paper, L is set to five, since five blocks are used in the PreAct ResNet-18 [10].

By combining the adversarial loss and attention transfer loss, the knowledge from the disturbance labeled face database is successfully transferred to the disturbance unlabeled FER database.

Attention Block. Inspired by [19], we develop an attention block. The network architecture of the attention block is given in Figure 3.

The first attention block in S_e or S_d takes the features $\mathbf{u}_1$ from the first convolution block in S_g as the input. For the subsequent attention block at the j-th layer, the element-wise weighted addition between the global shared features $\mathbf{u}_j$ in S_g and the task-specific features $\mathbf{a}_{j-1}^t$ ($t \in \{e, d\}$) from the previous layer in S_t ($t \in \{e, d\}$), is taken as the input, as illustrated in Figure 2(b). Then, the attention mask $\mathbf{m}_j^t$ ($t \in \{e, d\}$) generated from the j-th layer in S_t ($t \in \{e, d\}$) is expressed as:

$$\mathbf{m}_j^t = \begin{cases} \mathrm{g}(\mathbf{u}_j), & j = 1, \\ \mathrm{g}(\delta_1 \mathbf{u}_j + \delta_2 \mathbf{a}_{j-1}^t), & j \geq 2, \end{cases} \tag{8}$$

where δ_1 and δ_2 are the learnable parameters that respectively determine the importance of global shared features $\mathbf{u}_j$ and task-specific features $\mathbf{a}_{j-1}^i$; $\mathrm{g}(\cdot)$ denotes the aggregation of a batch normalization layer, a parametric ReLU layer, a 1×1 convolutional layer, another batch normalization layer, and a sigmoid layer that constrains the output within the range of $(0, 1)$.

The output feature maps of the j-th attention block for S_t ($t \in \{e, d\}$) are given as:

$$\mathbf{a}_j^t = \mathrm{h}(\mathbf{m}_j^t \odot \mathbf{u}_j), \tag{9}$$

where $\odot$ denotes the element-wise multiplication; $\mathrm{h}(\cdot)$ denotes a 3×3 convolutional layer to match the channels between the task-specific features in the j-th layer and the global shared features in the $(j+1)$-th layer, followed by a batch normalization layer, a parametric ReLU layer, and a max pooling layer to match the sizes of the feature maps between two types of features.

Multi-level Attention Mechanism. The features from different levels of the network are complementary. An example of the output feature maps in S_e is illustrated in Figure 4. We can see that high-level features extracted from deep layers are beneficial to locate salient regions, while low-level features extracted from shallow layers can be used to determine salient boundaries.

Based on the above observations, a multi-level attention mechanism is introduced in S_e. To be specific, we combine the output feature maps from different layers of S_e by cross-channel concatenation. Considering that the sizes of feature maps vary from layer to layer, we utilize several max pooling layers to ensure the same size of the feature maps from different attention blocks (except for the last block). Then, these resized feature maps are concatenated as:

$$\mathbf{a}_{out} = [\hat{\mathbf{a}}_1^e; \cdots; \mathbf{a}_L^e], \tag{10}$$

where $\hat{\mathbf{a}}_j^e$ is the output feature maps of the corresponding max pooling layer for $\mathbf{a}_j^e$. $\mathbf{a}_{out}$ is the final combined feature maps. In this way, both low-level spatial features and high-level semantic features are aggregated to extract expression-specific features.

Joint Loss Function. The joint loss function for the DDM is defined as:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{AD} + \lambda_2 \mathcal{L}_{AT}, \tag{11}$$

where λ_1 and λ_2 denote the weights of the adversarial loss and attention transfer loss, respectively.

By minimizing the joint loss function, the DDM is able to extract discriminative expression-specific features.

3.4 Discussions

A number of existing CNN based FER methods [23, 36] suffer from the problem that the final expression-specific features contain the disturbance information because of limited training data. Some

disturbance-disentangled based FER methods [22, 39] may not accurately recognize expressions in the disturbance unlabeled FER databases. In contrast, in our method, the two task-specific subnetworks are learned in a collaborative way. By explicitly designing a disturbance sub-network, the disturbance information can be effectively disentangled from the features used for expression recognition. Such a manner significantly improves the discriminability of expression-specific features. Meanwhile, based on adversarial transfer learning, the knowledge in the DFEM learned from the large-scale face database can be successfully transferred to the DDM to perform expression recognition on the disturbance unlabeled FER database. Therefore, the problems due to limited training data and the lack of disturbance label information can be greatly alleviated.

4 EXPERIMENTS

In this section, extensive experiments are performed to show the effectiveness of our proposed method. We first introduce the databases. Then, we show the implementation details. Next, we conduct the ablation studies to evaluate each component of our proposed method. Finally, we compare our method with several state-of-the-art FER methods.

4.1 Databases

CK+: The Extended Cohn-Kanade (CK+) database contains 327 video sequences annotated with expression labels, including six basic expressions (i.e., angry, happy, surprise, sad, disgust, fear) and one non-basic expression (contempt). We choose the three peak expressional frames from each sequence to construct the training set and the test set. We employ the popular ten-fold cross-validation protocol for evaluation in this paper, as done in [6, 22, 35, 44].

MMI: The MMI database is composed of 205 image sequences captured in the frontal view, and the sequences are labeled with six basic facial expressions. Similar to the CK+ database, we select the three peak expressional frames in each sequence for training and testing. The subject-independent ten-fold cross-validation is also conducted.

Oulu-CASIA: The Oulu-CASIA database contains videos of 80 subjects and six basic expressions. The images are captured with two imaging systems (i.e., near-infrared and visible light), under three different illumination conditions, including normal indoor illumination, weak illumination, and dark illumination. As done in [35], the last three frames in each sequence captured with the visible light and strong illumination are used in our experiments. The subject-independent ten-fold cross-validation is conducted.

RAF-DB: The real-world affective face database (RAF-DB) is a real-world database that contains 15,331 images labeled with six basic facial expressions and a neutral expression, where 12,271 and 3,068 images are used for training and testing, respectively. In addition to the expression label, the images in RAF are also labeled with the facial attributes of age, gender, and race.

SFEW: The SFEW database is created by selecting the static frames from the AFEW database [4]. It is very challenging because it covers unconstrained facial expressions, varied head poses, large age range, varied focus, different resolutions of faces and real-world illumination. Each image is assigned to one of six basic expression or the neutral expression. The training set contains 847 images and the test set contains 409 images.

Table 1: The details of the three baseline methods and six DDL variants.

Methods	S_g	S_e		S_d					
		w/o	multi	gen	age	race	id	ill	pose
Baseline	√	-	-	-	-	-	-	-	-
Baseline_at	√	√	-	-	-	-	-	-	-
Baseline_mat	√	-	√	-	-	-	-	-	-
DDL_g	√	-	√	√	-	-	-	-	-
DDL_ga	√	-	√	√	√	-	-	-	-
DDL_gar	√	-	√	√	√	√	-	-	-
DDL_gar&id	√	-	√	√	√	√	√	-	-
DDL_gar&id&il	√	-	√	√	√	√	√	√	-
DDL_gar&id&il&p	√	-	√	√	√	√	√	√	√

"w/o" and "multi" respectively represent that S_e is trained without and with the multi-level attention mechanism; "gen", "age", "race", "id", "ill", and "pose" denote that the DFEM is pre-trained to predict gender, age, race, identity, illumination, and pose, respectively.

4.2 Implementation Details

For all the databases, the face in each image is detected and cropped according to the eye positions. Then, the facial image is resized to the size of 100×100. During training, the facial images are randomly cropped to the size of 90×90, and the cropped images are further processed by using the horizontal flip. By default, the PreAct ResNet-18 is pre-trained on the AffectNet database [24].

The values of λ_1 and λ_2 in Eq. (11) are empirically set to 1.0 and 0.05, respectively. The dimension of disturbance-specific features is 128 and that of expression-specific features is equal to the number of expression categories (6 or 7 in the FER database). We train the networks using the Adam algorithm [14] with the initial learning rate of 0.0001, $\beta_1 = 0.500$, and $\beta_2 = 0.999$. The learning rate is further divided by 10 after 10, 18, 25, and 32 epochs. All the models are trained on a single NVIDIA GTX 1080 Ti using Pytorch for 40 epochs with a batch size of 16 for RAF-DB and 8 for the other FER databases.

The DFEM is pre-trained on the Multi-PIE face database [8] to extract features for classifying identity, pose, and illumination, where Multi-PIE contains 755,370 images from 337 subjects under 15 viewpoints and 20 light conditions. The DFEM is also pre-trained on the RAF-DB database to extract features for classifying gender, age, and race. When all the six disturbing factors are considered, the DFEM is pre-trained by combining Multi-PIE and RAF-DB, where missing labels are ignored during back-propagation.

4.3 Ablation Studies

In order to show the effectiveness of the proposed DDL method, we conduct the ablation studies to evaluate the influence of the attention block, the multi-level attention mechanism, and different disturbing factors on the performance of our proposed method. Specifically, we evaluate the performance of three baseline methods and six DDL variants, whose details are summarized in Table 1.

Table 2: The recognition accuracy (%) obtained by three baselines and six DDL variants. The best results are boldfaced.

Databases	Baseline	Baseline_at	Baseline_mat	DDL					
				g	ga	gar	gar&id	gar&id&il	gar&id&il&p
CK+	96.37	97.93	98.08	98.19	98.41	98.96	99.10	**99.16**	98.78
MMI	77.68	79.23	79.43	82.05	82.72	83.17	83.19	**83.67**	83.01
Oulu-CASIA	83.40	86.18	86.53	87.22	87.36	87.78	**88.26**	87.85	87.64
RAF-DB	85.89	86.63	86.90	87.19	87.22	87.45	87.45	87.55	**87.71**
SFEW	55.96	56.42	57.34	58.03	58.26	58.49	59.60	59.63	**59.86**

Figure 5: Visualization of attentive feature maps on the CK+, MMI, Oulu-CASIA, RAF-DB, and SFEW databases. Left to right in each panel: angry, surprise, disgust, fear, happy, and sad.

Table 2 shows the recognition accuracy comparison obtained by these nine methods.

Influence of Attention Block and Multi-level Attention Mechanism. As shown in Table 2, compared with Baseline, Baseline_at achieves 1.56%, 1.55% and 2.78% gains in terms of recognition accuracy on the CK+, MMI, and Oulu-CASIA databases, respectively. For the in-the-wild databases, its accuracy is improved by 0.74% and 0.46% on the RAF-DB and SFEW databases, respectively. These results demonstrate the effectiveness of the attention block. Furthermore, Baseline_mat achieves higher recognition accuracy than Baseline_at. Specifically, compared with Baseline_at, the recognition accuracy of Baseline_mat is improved by 0.15%, 0.20%, 0.35%, 0.27%, 0.92% on CK+, MMI, Oulu-CASIA, RAF-DB, and SFEW, respectively. This verifies the effectiveness of the multi-level attention mechanism.

To show the importance of the multi-level attention mechanism, we add the generated feature maps in S_e to the input facial images and visualize them in Figure 5. To be specific, the combined feature maps (see Eq. (10)) with a size of $1472 \times 6 \times 6$ before the fully-connected layer are first added along the channel dimension, which generates an attentive feature map with the size of 6×6. Then, this feature map is resized to the same size as the input image. Finally, we add the resized attentive feature map to the input image and obtain the final results. As shown in Figure 5, the warm-toned parts of an image correspond to the regions with large values in the attentive feature map, and vice versa. We can observe that the attentive feature map is able to focus on key facial regions (especially the regions around eyes and mouth) that are critical for expression recognition. In particular, for the images in the in-the-lab databases, the corresponding attentive feature maps focus on small facial patches. For the images in the in-the-wild databases, the corresponding attentive feature maps tend to pay attention to relatively large facial patches. This is because that the images in the in-the-wild databases involve large pose variations and low image quality. A larger facial patch is beneficial to extract more discriminative features for predicting facial expressions on the in-the-wild databases.

Influence of Different Disturbing Factors. The influence of different disturbing factors on FER is also shown in Table 2.

We can observe that all the variants consistently perform better than Baseline_mat, which demonstrates the importance of the disturbance sub-network S_d. S_d is helpful to disentangle the disturbance information from facial expression images, and enable the model to extract highly discriminative expression-specific features, thus improving the final performance.

In addition, for the in-the-wild databases, the recognition accuracy obtained by DDL tends to be higher when more disturbing factors are considered. The proposed method achieves the best performance when all the disturbing factors are employed in the DFEM. This is because that the images in the in-the-wild databases usually contain severe variations caused by multiple disturbing factors. Explicitly disentangling these disturbing information has a positive influence on the extraction of effective expression-specific features. However, the proposed method obtains the top accuracy on the CK+ and MMI databases when all the disturbing factors except for the pose are considered. This is because the images in the in-the-lab database are all frontal images. As a result, considering the pose as the disturbing factor in the DFEM leads to performance decrease. Meanwhile, the proposed method performs best on the Oulu-CASIA database, when all the disturbing factors except for the pose and illumination are considered. This can be ascribed to the fact that the images in Oulu-CASIA do not contain obvious pose and illumination variations. Therefore, it is critical to properly choose the disturbing factors by taking into account the characteristics of the FER database.

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

In this subsection, we compare the proposed method with several state-of-the-art FER methods. Table 3 and Table 4 give the performance obtained by all the competing methods on the in-the-lab databases and the in-the-wild databases, respectively.

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

Methods	Accuracy (%)		
	CK+	MMI	Oulu-CASIA
LBP-TOP [42]	88.99‡	59.51	68.13
PPDN [44]	97.30†	-	72.40
IACNN [22]	95.37‡	71.55	-
DLP-CNN [16]	95.78†	78.46	-
DTAGN* [13]	97.25‡	70.20	81.46
IPA2LT [37]	92.45‡	65.61	61.49
DAM-CNN [32]	95.88†	-	-
L2-sparseness[33]	97.59‡	78.54	82.92
DeRL [35]	97.37‡	73.23	88.00
PHRNN-MSCNN* [40]	98.50‡	81.18	86.25
FN2EN [6]	98.60†	-	87.71
DDL (proposed)	**99.16‡**	**83.67**	**88.26**

‡ and † respectively denote that seven expression categories and six expression categories are used in CK+; * indicates that the method is trained based on the image sequences.

As shown in Table 3, we can observe that almost all the methods obtain high recognition accuracy on the CK+ database, while achieving relatively worse performance on the MMI and Oulu-CASIA databases. This is because that the images from CK+ are of high quality and the intensities of different expressions are relatively strong, while those from MMI are affected by the glasses and the expression intensities from Oulu-CASIA are relatively weak.

Among all the competing methods, the top three methods are our proposed DDL, FN2EN [6], and PHRNN-MSCNN [40]. Our proposed DDL achieves better performance than FN2EN, even though our test set is more challenging. The FN2EN method only classifies the basic six expression categories in CK+. On the contrary, our method classifies not only the basic six expressions, but also the contempt category in CK+. The recurrent neural network (RNN) is used in PHRNN-MSCNN, where both the facial image and the facial landmarks are used as the inputs. In contrast, our proposed method only uses a single image as the input. Nevertheless, our proposed DDL still achieves the best performance among all the competing methods, which can be ascribed to the effectiveness of our proposed deep disturbance-disentangled learning and the multi-level attention mechanism.

As shown in Table 4, we compare our method with eight state-of-the-art methods on the in-the-wild databases. Among them, DLP-CNN [17] proposes a locality preserving loss to implicitly address the disturbance problem and reduce the intra-class distance. IPA2LT [37] addresses the problem of inconsistent annotations in the FER databases. SPDNet [1] introduces the covariance pooling into FER and achieves state-of-the-art performance. However, the above methods do not explicitly deal with the disturbing factors, thus leading to inferior performance.

IACNN [22] proposes an identity-aware network by taking identity into account. IPFR [29] simultaneously considers identity and pose in their framework. These methods can only address one or two disturbing factors, whose labels are given in the FER databases. Different from the above methods, our proposed method is able to explicitly disentangle multiple disturbing factors by leveraging adversarial transfer learning, even though disturbing factors are not labeled in the FER databases. gACNN [18] and RAN [30] explicitly handle the occlusion problem through combining local learning and global learning. However, these two methods only utilize the high-level features to perform FER. Unlike these two methods, our proposed method exploits both high-level features and low-level features, thereby achieving the best performance. This is due to the fact that we take advantage of the large-scale disturbance labeled face database to explicitly learn the distribution of multiple disturbing factors, and entangle the disturbance by elaborately designing an expression sub-network and a disturbance sub-network based on a global shared sub-network.

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

Methods	Accuracy (%)	
	RAF-DB	SFEW
IACNN [22]	-	50.98
DLP-CNN [17]	84.13	51.05
gACNN [18]	85.07	-
IPA2LT [37]	86.77	58.29
SPDNet [1]	87.00	58.14
IPFR [29]	-	57.40
DAM-CNN [32]	-	42.30
RAN [30]	86.90	56.40
DDL (proposed)	**87.71**	**59.86**

5 CONCLUSIONS

In this paper, we propose a novel DDL method for FER. DDL is able to simultaneously disentangle multiple disturbing factors (even when the labels of these disturbing factors are not available in the FER databases) and effectively extract the expression-related information, by taking advantage of multi-task learning and adversarial transfer learning. The training of DDL contains two stages: pre-train a DFEM to extract features for classifying multiple disturbing factors, and train a DDM to extract expression-specific features and disturbance-specific features. In particular, a multi-level attention mechanism is employed in the expression-specific sub-network to make full use of both low-level and high-level features. Extensive experiments conducted on three in-the-lab FER databases and two in-the-wild FER databases have demonstrated the superior performance of our proposed method.

Currently, our method cannot adaptively choose the disturbing factors when evaluated on a FER database. In the future, we intend to investigate effective ways to adaptively disentangle appropriate disturbing factors on different FER databases.

ACKNOWLEDGMENTS

This work was partly supported by the National Key R&D Program of China under Grant 2017YFB1302400 and by the National Natural Science Foundation of China under Grants U1605252 and 61872307.

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] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L. Yuille, and Xiaogang Wang. 2017. Multi-context attention for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1831–1840.

[3] Charles Darwin and Phillip Prodger. 1998. *The expression of the emotions in man and animals*. Oxford University Press, USA.

[4] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia* 3 (2012), 34–41.

[5] 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.

[6] 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.

[7] Jianlong Fu, Heliang Zheng, and Tao Mei. 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4438–4446.

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

[9] 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.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision*. 630–645.

[11] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7132–7141.

[12] Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao, and Yurong Chen. 2017. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the ACM International Conference on Multimedia Interaction*. 553–560.

[13] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. 2015. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 2983–2991.

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

[15] Shan Li and Weihong Deng. 2018. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348* (2018).

[16] 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.

[17] 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.

[18] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2018. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing* 28, 5 (2018), 2439–2450.

[19] Shikun Liu, Edward Johns, and Andrew J Davison. 2019. End-to-end multi-task learning with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1871–1880.

[20] 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.

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

[22] 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.

[23] Ali Mollahosseini, David Chan, and Mohammad H Mahoor. 2016. Going deeper in facial expression recognition using deep neural networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. 1–10.

[24] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (2017), 18–31.

[25] Bowen Pan, Shangfei Wang, and Bin Xia. 2019. Occluded facial expression recognition enhanced through privileged information. In *Proceedings of the ACM International Conference on Multimedia*. 566–573.

[26] Maja Pantic and Leon J. M. Rothkrantz. 2000. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 12 (2000), 1424–1445.

[27] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[28] 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.

[29] 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.

[30] 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, (2020), 4057–4069.

[31] Lin Wu, Yang Wang, Junbin Gao, and Xue Li. 2018. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. *IEEE Transactions on Multimedia* 21, 6 (2018), 1412–1424.

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

[33] Weicheng Xie, Xi Jia, Linlin Shen, and Meng Yang. 2019. Sparse deep feature learning for facial expression recognition. *Pattern Recognition* 96 (2019), 1–13.

[34] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*. 2048–2057.

[35] 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.

[36] Zhiding Yu and Cha Zhang. 2015. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the ACM on International Conference on Multimedia Interaction*. 435–442.

[37] 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.

[38] 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.

[39] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. 2018. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3359–3368.

[40] 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.

[41] 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.

[42] Guoying Zhao and Matti Pietikainen. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 6 (2007), 915–928.

[43] Sicheng Zhao, Zizhou Jia, Hui Chen, Leida Li, Guiguang Ding, and Kurt Keutzer. 2019. PDANet: Polarity-consistent deep attention network for fine-grained visual emotion regression. In *Proceedings of the ACM International Conference on Multimedia*. 192–201.

[44] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. 2016. Peak-piloted deep network for facial expression recognition. In *Proceedings of the European Conference on Computer Vision*. 425–442.