International Journal of Computer Vision
https://doi.org/10.1007/s11263-021-01556-7

Adaptive Deep Disturbance-Disentangled Learning for Facial Expression Recognition

Delian Ruan[1] · Rongyun Mo[1] · Yan Yan[1] · Si Chen[2] · Jing-Hao Xue[3] · Hanzi Wang[1]

Received: 22 December 2020 / Accepted: 24 November 2021

Abstract
In this paper, we propose a novel adaptive deep disturbance-disentangled learning (ADDL) method for effective facial expression recognition (FER). ADDL involves a two-stage learning procedure. First, a disturbance feature extraction model is trained to identify multiple disturbing factors on a large-scale face database involving disturbance label information. Second, an adaptive disturbance-disentangled model, which contains a global shared subnetwork and two task-specific subnetworks, is designed and learned to explicitly disentangle disturbing factors from facial expression images. In particular, the expression subnetwork leverages a multi-level attention mechanism to extract expression-specific features, while the disturbance subnetwork embraces a new adaptive disturbance feature learning module to extract disturbance-specific features based on adversarial transfer learning. Moreover, a mutual information neural estimator is adopted to minimize the correlation between expression-specific and disturbance-specific features. Extensive experimental results on both in-the-lab FER databases (including CK+, MMI, and Oulu-CASIA) and in-the-wild FER databases (including RAF-DB, SFEW, Aff-Wild2, and AffectNet) show that our proposed method consistently outperforms several state-of-the-art FER methods. This clearly demonstrates the great potential of disturbance disentanglement for FER. Our code is available at https://github.com/delian11/ADDL.

Keywords Facial expression recognition · Multi-task learning · Adversarial transfer learning · Multi-level attention

1 Introduction

Facial expression conveys nonverbal cues and plays a fundamental role in understanding emotions in human interaction and communication. During the past few decades, facial expression recognition (FER) has attracted increasing attention in computer vision due to its variety of applications in entertainment, sociable robotics, data-driven animation, and so on (Zhang et al., 2018a, b). Recently, with the considerable development of deep learning, FER has made substantial progress (Chang et al., 2019; Chen et al., 2020; Dapogny et al., 2018; Kollias et al., 2020a; Li & Deng, 2019; Li et al., 2017; Meng et al., 2017; Yang et al., 2018a; Yan et al., 2020; Zhang et al., 2018c).

Despite great progress, FER is still a challenging task. On the one hand, facial expression images exhibit large inter-class similarities and intra-class variances caused by the existence of multiple disturbing factors. For example, in

Communicated by Maja Pantic.

Delian Ruan and Rongyun Mo have contributed equally to this work.

✉ Yan Yan
yanyan@xmu.edu.cn

Delian Ruan
delianruan@stu.xmu.edu.cn

Rongyun Mo
morongyun@stu.xmu.edu.cn

Si Chen
chensi@xmut.edu.cn

Jing-Hao Xue
jinghao.xue@ucl.ac.uk

Hanzi Wang
hanzi.wang@xmu.edu.cn

[1] Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China

[2] School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China

[3] Department of Statistical Science, University College London, London, WC1E 6BT, UK

Published online: 05 January 2022

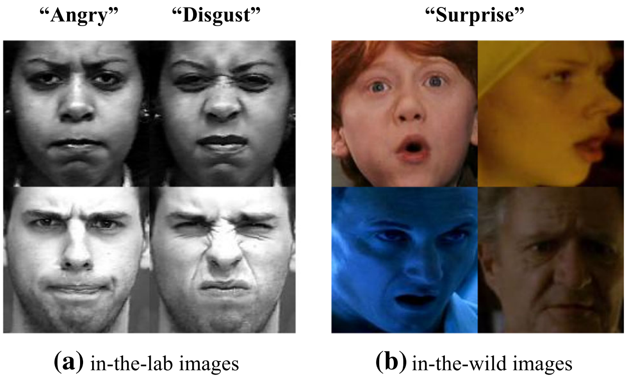


Fig. 1 Some facial expressions of **a** in-the-lab images [the images are from the CK+ database (Lucey et al., 2010)] and **b** in-the-wild images [the images are from the SFEW database (Dhall et al., 2011)]

each row of Fig. 1a, the images of different expressions are visually similar due to the same illumination and identity. Meanwhile, in the two rows of Fig. 1b, the images of the same expression show great differences because of changes in gender, age, race, identity, illumination, and pose. Clearly, these disturbing factors adversely affect the extraction of expression-specific features. On the other hand, different FER databases may involve different types of disturbing factors. For instance, some in-the-lab FER databases only include disturbances caused by variations in identity, age, and gender but not pose, as shown in Fig. 1a, while some in-the-wild FER databases may suffer from severe identity, illumination, and pose variations, as given in Fig. 1b.

It is of great importance to properly disentangle disturbing factors from facial expression images for FER. A variety of deep learning-based FER methods (Rifai et al., 2012; Liu et al., 2018; Mollahosseini et al. 2016; Hu et al., 2017; Wang et al., 2020c) have been proposed to implicitly reduce the disturbance for recognizing facial expressions. The training of these methods typically relies on a large amount of labeled data to achieve satisfactory performance. However, many FER databases contain only limited labeled training data. As a result, it is not a trivial task to learn robust deep models that can effectively alleviate the influence of various disturbing factors in the case of limited training data.

To date, some disturbance-disentangled-based FER methods (Meng et al., 2017; Zhang et al., 2020b; Chen et al., 2018; Zhang et al., 2018b; Yang et al., 2018b), which explicitly disentangle disturbing factors, have been developed. Note that many FER databases only provide labels of facial expression and identity (or pose) since manually labeling various disturbing factors is time-consuming and labor-intensive. As a consequence, these methods are only able to disentangle one or two disturbing factors for FER, leading to suboptimal performance. Moreover, they may not work well when the labels of disturbing factors are not available in the FER databases.

Fortunately, some large-scale face databases provide a large number of facial images together with the label information for different disturbing factors. For example, Multi-PIE (Gross et al., 2010) offers labels of identity, pose, and illumination. RAF-DB (Li et al., 2017) gives labels of gender, race, and age. Therefore, how to effectively exploit these large-scale disturbance-labeled face databases to perform transfer learning for classifying expressions in disturbance-unlabeled FER databases is a significantly rewarding research problem.

To address the above problems, we propose a novel adaptive deep disturbance-disentangled learning (ADDL) method for FER. ADDL adaptively disentangles multiple disturbing factors from facial expression images and effectively extracts expression-specific features, building its success by borrowing the strengths from both multi-task learning and adversarial transfer learning.

The ADDL method involves a two-stage learning procedure: (1) training a disturbance feature extraction model (DFEM) and (2) training an adaptive disturbance-disentangled model (ADDM). Specifically, the DFEM is first trained to identify multiple disturbing factors on a large-scale face database. Second, based on the trained DFEM, the ADDM is learned to remove the disturbance and extract discriminative features for expression recognition.

In summary, the main contributions of our work are as follows:

- We propose a novel ADDL method that contains two crucial components (i.e., the DFEM and ADDM) for effective FER. In particular, the knowledge in the DFEM trained on the large-scale face database is effectively transferred to the ADDM to classify expressions in the disturbance-unlabeled FER database. Therefore, the ADDL method is capable of simultaneously disentangling multiple disturbing factors and capturing expression-related information.
- We elaborately design two task-specific subnetworks in the ADDM. For the expression subnetwork, we employ a multi-level attention mechanism to extract expression-specific features. For the disturbance subnetwork, we adopt adversarial transfer learning to learn disturbance-specific features. The two subnetworks are jointly trained to exploit both spatial-aware and semantic-aware information.
- We extensively evaluate the proposed ADDL on both in-the-lab and in-the-wild FER databases. Experimental results from these databases show that our proposed method performs favorably against several state-of-the-art FER methods.

This paper is a substantial extension of our previous conference work in Ruan et al. (2020). The method in our previous work has two main limitations. First, it cannot adaptively choose the disturbing factors when trained on an FER database. Second, disturbance disentanglement is not explicitly performed. This paper alleviates these limitations in two aspects: (1) an adaptive disturbance feature learning module (ADFL) is designed to learn the importance weights corresponding to different disturbing factors and then perform adversarial transfer learning; (2) a mutual information neural estimator (MINE) is leveraged to minimize the correlation between expression-specific and disturbance-specific features.

To summarize, we have added the following new significant contributions:

- We design the ADFL to greatly facilitate the extraction of disturbance-specific features by considering different influences of disturbing factors in the FER training database. In this way, the characteristics of the FER database can be taken into account to choose the disturbing factors, and thus adaptive disturbance-specific features are extracted.
- We adopt the MINE during the training of the ADDM. Thus, we are able to effectively enhance the explicit disentanglement between expression-specific and disturbance-specific features for better FER.
- Based on the above two extensions, the proposed ADDL method consistently achieves better recognition accuracy than our previous method on both in-the-lab and in-the-wild FER databases.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related work. Section 3 introduces the details of our proposed ADDL method. Section 4 provides experimental results on three popular in-the-lab FER databases (CK+, MMI, and Oulu-CASIA) and four challenging in-the-wild FER databases (RAF-DB, SFEW, AffWild-2, and AffectNet). Finally, Sect. 5 presents the conclusion and future work.

2 Related Work

In this section, we review state-of-the-art work of convolutional neural network (CNN)-based FER methods in Sect. 2.1, disturbance-disentangled-based FER methods in Sect. 2.2, action unit recognition in Sect. 2.3, and attention mechanisms in Sect. 2.4, which are closely related to our proposed method.

2.1 CNN-Based FER Methods

Currently, CNN-based FER methods (Li & Deng, 2020) have achieved promising performance due to the powerful capability of CNNs to capture high-level semantic information. For example, Yu and Zhang (2015) develop an ensemble of CNNs, which shows impressive results in the EmotiW challenge. Mollahosseini et al. (2016) introduce a network consisting of two convolutional layers and four inception layers (Szegedy et al., 2015) to predict facial expressions. Hu et al. (2017) propose a supervised scoring ensemble (SSE) method, where supervision signals are used not only for deep layers but also for intermediate and shallow layers. Liu et al. (2018) design a multi-channel pose-aware CNN (MPCNN) to aggregate multi-scale features, which are fed into a pose-aware recognition network for pose estimation and pose conditioned expression recognition.

These CNN-based methods implicitly alleviate the influence of disturbances involved in facial expression images. Generally, they rely heavily on a large number of labeled data to learn effective feature representations. However, many FER databases do not provide sufficient training data containing diverse variations for different disturbing factors. As a result, the trained CNN models may not be sufficiently robust to handle various disturbing factors.

2.2 Disturbance-Disentangled-Based FER Methods

Some methods have been proposed to explicitly perform disturbance disentanglement for FER. For example, Meng et al. (2017) introduce an identity-aware CNN (IACNN) method to alleviate the variations caused by facial identity, where an identity-sensitive contrastive loss is developed to learn identity-related information. Wang et al. (2019) propose an adversarial feature learning method to disentangle the disturbance caused by pose and identity.

Recently, generative adversarial networks (GANs) have been widely used in pose-robust FER (Zhang et al., 2018b; Wang et al., 2020d) and identity-robust FER (Chen et al., 2018; Yang et al., 2018b). Zhang et al. (2018b, 2020a, 2020b) develop a GAN-based pose-invariant method for facial image synthesis and expression recognition by exploiting the relationship between different poses and expressions. Furthermore, the disturbance caused by facial identity is explicitly reduced by adversarial learning. Yang et al. (2018b) propose an identity-adaptive method to learn an identity subspace, which can generate different expressions while preserving identity-related information for each individual.

The above methods require the labels of disturbing factors in the FER training databases. Unfortunately, many FER databases only provide labels of facial expressions and some facial attributes (such as identity and pose) but lack the label information for other disturbing factors. Therefore, these methods are only able to handle one or two disturbing factors. Moreover, they may fail on disturbance-unlabeled FER databases.

Facial expression images are often intertwined with multiple disturbing factors (such as identity, pose, age, gender, and illumination). Hence, it is desirable to simultaneously alleviate the influence of these disturbing factors. In this paper, we capitalize on the disturbance label information available in the large-scale face database to perform adversarial transfer learning for expression recognition on the disturbance-unlabeled FER database. This not only successfully addresses the problems of the lack of disturbance labels and limited training data in the FER database, but also enables the proposed method to effectively disentangle different disturbing factors from facial expression images.

2.3 Action Unit Recognition

Ekman and Friesen (1976) develop the facial action coding system (FACS), which encodes atomic nonoverlapping facial muscles called action units (AUs). Based on the FACS, facial expressions can be viewed as combinations of certain AUs.

Some methods have been proposed to learn task-specific representations for AU recognition. For example, Zhang et al. (2018d) design an adversarial training framework (ATF), which is trained by minimizing the AU loss and maximizing the identity loss. In this way, identity-invariant features are extracted for AU detection. Li et al. (2019) propose a twin-cycle autoencoder (TCAE) for AU detection in a self-supervised manner. They factorize the movements into AU-related and pose-related displacements based on a pair of images. Therefore, facial action-related movements can be disentangled from head motion-related movements, which is beneficial for learning discriminative AU-related features. Sankaran et al. (2020) implicitly capture the correlations between two modalities by using an encoder-decoder framework to learn a unified representation for cross-modality AU recognition.

The above methods perform disentanglement based on multi-task learning CNN or an encoder-decoder structure. Nevertheless, they take one or two disturbing factors into account and do not fully consider the explicit disentanglement between the AU-related movements and disturbing factors.

2.4 Attention Mechanisms

In recent years, attention mechanism-based CNN methods have been developed in a variety of tasks, such as fine-grained image recognition (Fu et al., 2017; Hu et al., 2018), image captioning (Xu et al., 2015), person re-identification (Wu et al., 2018), and human pose estimation (Chu et al., 2017). Hu et al., (2018) propose a novel architecture unit termed the squeeze-and-excitation (SE) block, which adaptively recalibrates channel-wise feature responses by modeling interdependencies between channels. Chu et al. (2017) design a multi-context attention mechanism-based network for human pose estimation.

Psychological studies have revealed that salient facial regions (such as the mouth, nose, and eyes) play a critical role in FER (Pantic & Rothkrantz, 2000). Meanwhile, attention mechanisms have shown great capability to select salient features. Therefore, attention mechanisms are beneficial to improve the FER performance. For instance, Xie et al. (2019a) propose a deep attentive multi-path CNN (DAM-CNN) method for FER, where a spatial attention mechanism is adopted to obtain salient regions. Wang et al. (2020c) propose a region attention network (RAN) to locate salient facial regions for occlusion-invariant and pose-invariant FER. In general, these methods leverage high-level semantic features of CNNs for FER.

Both high-level and low-level features of CNNs are advantageous for performing FER. Low-level features capture the spatial-aware information of facial images, which can be used to determine the boundaries of salient regions. High-level features encode the semantic-aware information of facial images, which is desirable to locate salient regions (Zhao & Wu, 2019). In this paper, unlike previous methods, we employ a multi-level attention mechanism, which aggregates the attentive features from different layers of the network. This mechanism effectively exploits both spatial-aware and semantic-aware information to extract discriminative features for identifying facial expressions. Moreover, we leverage a self-attention layer to learn the importance weights corresponding to different disturbing factors, enabling the extraction of adaptive disturbance-specific features. Therefore, we can accommodate the different influences of multiple disturbing factors in the FER database.

3 Proposed Method

In this section, we introduce our proposed ADDL method in detail. First, an overview of the ADDL method is given in Sect. 3.1. Then, the key components (the DFEM and ADDM) of ADDL are described in Sects. 3.2 and 3.3, respectively. Finally, some discussions about ADDL are presented in Sect. 3.4.

3.1 Overview

The training phase of the ADDL method involves a two-stage learning procedure: (1) training a DFEM to predict various disturbing factors, and (2) training an ADDM, which adapts to the characteristics of each FER database, to extract expression-specific features by explicitly disentangling multiple disturbing factors from facial expression images. The network architecture of our proposed ADDL method is illustrated in Fig. 2.

Specifically, in the first stage, a DFEM is trained to simultaneously identify various disturbing factors on the

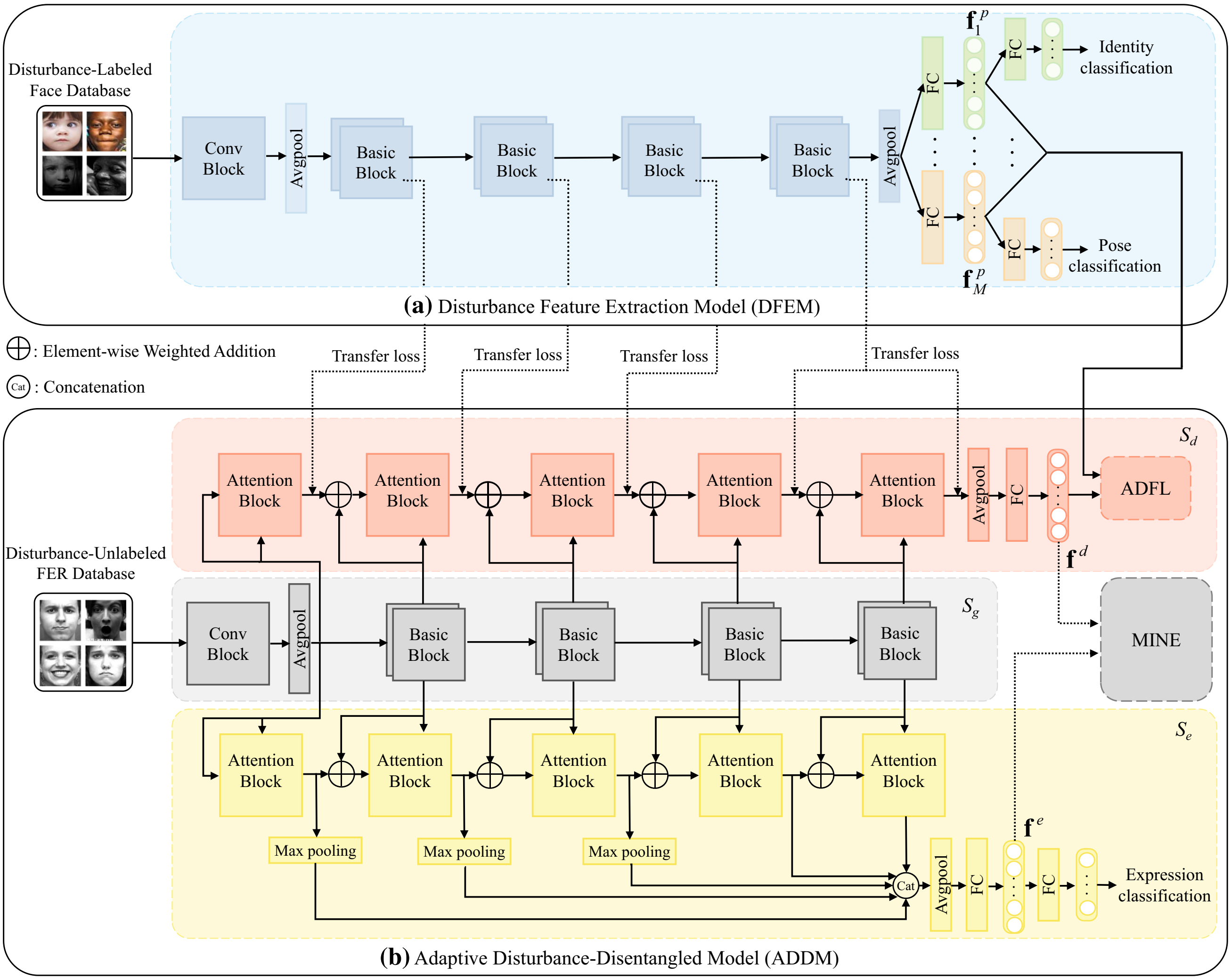


Fig. 2 The network architecture of our proposed ADDL method. The training phase of ADDL involves a two-stage learning procedure. **a** Training a DFEM consisting of shared layers and task-specific layers. The DFEM predicts various disturbing factors. **b** Training an ADDM consisting of a global shared subnetwork (S_g), an expression subnetwork (S_e), and a disturbance subnetwork (S_d). The ADDM extracts expression-specific features by explicitly disentangling the disturbance

disturbance-labeled face database. In this manner, the DFEM effectively captures the prior disturbance information. In the second stage, based on the trained DFEM, an ADDM, consisting of a global shared subnetwork and two task-specific subnetworks (i.e., an expression subnetwork and a disturbance subnetwork), is learned to classify expressions on the disturbance-unlabeled FER database.

In the ADDM, the expression subnetwork leverages a multi-level attention mechanism to comprehensively extract expression-specific features. Meanwhile, by taking advantage of adversarial transfer learning, the disturbance subnetwork capitalizes on the features extracted from the trained DFEM to effectively learn adaptive disturbance-specific features.

During the testing phase, given a facial image, only the global shared subnetwork and expression subnetwork from the trained ADDM are used to extract features and predict facial expressions.

3.2 Disturbance Feature Extraction Model (DFEM)

The DFEM is designed to extract discriminative features that capture the information for identifying multiple disturbing factors by taking advantage of multi-task learning on the disturbance-labeled face database. As shown in Fig. 2a, the network architecture of the DFEM consists of shared layers and task-specific layers.

Specifically, facial images are first fed into several shared layers consisting of a cascade of linear and nonlinear transformations to obtain high-level features. In this paper, we adopt ResNet-18 (He et al., 2016), which is widely used in previous works (Wang et al., 2020b), as shared layers. Then, the task-specific layers use a multi-branch architecture to extract features, where each branch containing two cascaded fully-connected (FC) layers classifies a disturbing factor. Note that the features obtained from the first FC layer of each branch encode the information for predicting a disturbing factor, while those from the second FC layer are the predicted outputs.

Given a disturbance-labeled face database, its training set $\mathbf{T}^l$ with R images is represented as $\mathbf{T}^l = \{\mathbf{x}_i^l, \mathbf{y}_i\}_{i=1}^R$, where $\mathbf{x}_i^l$ denotes the ith training image and $\mathbf{y}_i = [y_i^1, \ldots, y_i^M]^{\mathrm{T}}$ is an M-dimensional vector representing the labels of disturbing factors corresponding to $\mathbf{x}_i^l$. M denotes the number of disturbing factors. The optimization problem of the DFEM is formulated 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 features for the jth disturbing factor; $\mathcal{F}_j(\cdot, \cdot, \cdot)$ represents the prediction function for the jth disturbing factor, given the input $\mathbf{x}_i^l$ and the network parameters $\mathbf{w}_c$ and $\mathbf{w}_j$; y_i^j denotes the label of the jth disturbing factor corresponding to $\mathbf{x}_i^l$; and $\mathcal{L}_{CE}^j(\cdot, \cdot)$ represents the cross-entropy (CE) loss between the ground-truth label y_i^j and the result estimated by $\mathcal{F}_j$. Mathematically, the CE loss is defined as

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

where $\log(\cdot)$ represents the logarithm function; K^j indicates the category number of the jth disturbing factor; and $\mathbb{1}_{[k=y_i^j]}$ outputs 1 when $k = y_i^j$ and 0 otherwise.

3.3 Adaptive Disturbance-Disentangled Model (ADDM)

Based on the DFEM trained on the large-scale face database, the ADDM is learned to model the expression-related information and disturbance-related information on the disturbance-unlabeled FER database. As shown in Fig. 2b, the network architecture of the ADDM consists of a global shared subnetwork, an expression subnetwork, and a disturbance subnetwork.

In the following, we introduce the key components of the ADDM.

3.3.1 Global Shared Subnetwork

The global shared subnetwork (denoted S_g) is designed to extract global shared features of input images. In this paper, we employ ResNet-18 (He et al., 2016) as S_g, where the final FC layer is removed.

3.3.2 Task-Specific Subnetworks

ADDM contains two task-specific subnetworks, i.e., an expression subnetwork (denoted S_e) and a disturbance subnetwork (denoted S_d). Two subnetworks are jointly trained based on S_g.

Expression Subnetwork S_e is designed to learn expression-specific features by applying attention blocks to S_g. S_e consists of a set of attention blocks (see Sect. 3.3.3), which are followed by an average pooling layer and two FC layers. Here, the attention block generates a soft attention mask, which indicates the importance of each position in the feature map from S_g.

Considering that the features from different levels of the network in S_e are complementary, a multi-level attention mechanism is employed to fully exploit these features. Specifically, we first utilize several max pooling layers to ensure the same sizes of feature maps from different attention blocks (except for the last two blocks) since the sizes of feature maps vary from layer to layer. Then, these resized feature maps are concatenated as

$$\mathbf{a}_{out} = [\hat{\mathbf{a}}_1^e; \ldots; \hat{\mathbf{a}}_{L-2}^e; \mathbf{a}_{L-1}^e; \mathbf{a}_L^e], \tag{3}$$

where $\mathbf{a}_j^e$ indicates the feature map from the jth attention block in S_e; $\hat{\mathbf{a}}_j^e$ represents the output feature map of the max pooling layer corresponding to $\mathbf{a}_j^e$; L denotes the number of attention blocks; and $\mathbf{a}_{out}$ is the final combined feature map.

Note that, as shown in Fig. 2, the max pooling layer is not applied to $\mathbf{a}_{L-1}^e$ and $\mathbf{a}_L^e$ to ensure the same sizes of feature maps for concatenation. In this way, both low-level spatial features and high-level semantic features are aggregated to extract expression-specific features.

Given a disturbance-unlabeled FER database, its training set $\mathbf{T}^u$ with N images is represented as $\mathbf{T}^u = \{\mathbf{x}_i^u, y_i\}_{i=1}^N$, where $\mathbf{x}_i^u$ denotes the ith training image and y_i indicates the expression label corresponding to $\mathbf{x}_i^u$. S_e optimizes the following problem:

$$\underset{\mathbf{w}_g, \mathbf{w}_e}{\arg\min} \sum_{i=1}^{N} \mathcal{L}_{CE}(y_i, \mathcal{F}_e(\mathbf{x}_i^u, \mathbf{w}_g, \mathbf{w}_e)), \tag{4}$$

where $\mathbf{w}_g$ and $\mathbf{w}_e$ denote the network parameters in S_g and S_e, respectively; $\mathcal{F}_e$ denotes the prediction function; and $\mathcal{L}_{CE}$

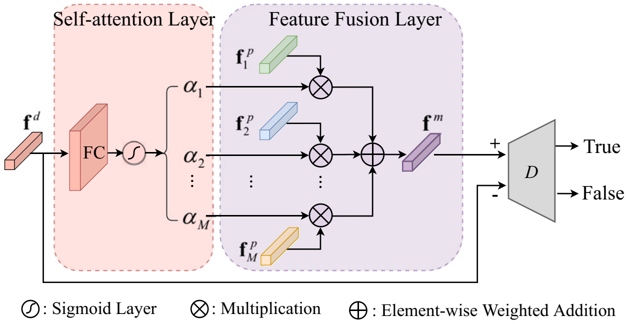


Fig. 3 The network architecture of the ADFL

indicates the CE loss between the ground-truth expression label y_i and the predicted result by $\mathcal{F}_e$, which is expressed as

$$\mathcal{L}_{CE} = -\sum_{k=1}^{K} \mathbb{1}_{[k=y_i]} \log(\mathcal{F}_e(\mathbf{x}_i^u, \mathbf{w}_g, \mathbf{w}_e)), \tag{5}$$

where K is the number of expression categories.

Disturbance Subnetwork S_d is designed to learn disturbance-specific features. To achieve this, a straightforward way is to generate pseudo-labels of disturbing factors in the FER database by applying the trained DFEM and then train S_d with these pseudo-labels. However, these pseudo-labels unavoidably involve a large number of noisy labels due to the discrepancy between the source domain (the face database used to train the DFEM) and the target domain (the FER database used to train S_d). As a result, these noisy labels seriously affect the extraction of disturbance-specific features, thereby reducing the final FER performance.

In this paper, we take advantage of adversarial transfer learning to effectively improve the performance of the model in the unlabeled target domain, given the labeled source domain. Mathematically, we constrain the distributions of the output feature maps from S_d to be as close as those from the trained DFEM. Such a manner alleviates the domain discrepancy and avoids labeling disturbing factors in the target domain.

The network architecture of S_d is given in Fig. 2b. It is comprised of a set of attention blocks, which are followed by an average pooling layer, an FC layer, and an adaptive disturbance feature learning module (ADFL), where the FC layer extracts disturbance-specific features and the ADFL performs adversarial transfer learning between disturbance-specific features and weighted disturbing factor features.

The network architecture of the ADFL is shown in Fig. 3. The ADFL consists of a self-attention (SA) layer (including an FC layer and a sigmoid layer), a feature fusion layer, and a discriminator.

We suppose that the extracted disturbance-specific feature is denoted $\mathbf{f}^d$, given a facial image from the disturbance-unlabeled FER database. First, the SA layer outputs the importance weights (represented as $[\alpha_1, \cdots, \alpha_M]^{\mathrm{T}}$) corresponding to M disturbing factors. These importance weights reflect the different influences of disturbing factors in the FER training database. Meanwhile, we also obtain a set of disturbing factor features extracted from the first FC layers of the task-specific layers in the trained DFEM, denoted $\mathbf{T}^p = \{\mathbf{f}_j^p\}_{j=1}^M$. Here, $\mathbf{f}_j^p$ represents the jth disturbing factor feature.

Then, the feature fusion layer combines these disturbing factor features according to their corresponding importance weights, which can be expressed as

$$\mathbf{f}^m = \sum_{j=1}^{M} \alpha_j \mathbf{f}_j^p, \tag{6}$$

where $\mathbf{f}^m$ represents the weighted disturbing factor feature.

Finally, a discriminator D (consisting of four FC layers and a leaky ReLU function) is introduced to play an adversarial game with a feature extractor F. Here, the feature extractor F refers to the layers used to extract $\mathbf{f}^d$ in S_d. F tries to minimize the divergence of the feature distributions between $\mathbf{f}^d$ and $\mathbf{f}^m$, while D aims to distinguish $\mathbf{f}^d$ from $\mathbf{f}^m$. The objective of adversarial training is formulated as

$$\min_D \max_F \mathcal{L}_{AD}(F, D), \tag{7}$$

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

$$\mathcal{L}_{AD} = -\mathbb{E}[\log(D(\mathbf{f}^m))] - \mathbb{E}[\log(1 - D(\mathbf{f}^d))]. \tag{8}$$

To facilitate knowledge transfer from the trained DFEM to S_d, it is natural that the distributions of both the final output features and the intermediate attention maps from S_d are close to those from the trained DFEM. Therefore, we also apply attention transfer (Zhang et al., 2018a), 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

$$\mathcal{L}_{AT} = \sum_{j=1}^{L} || \frac{\mathbf{q}_j^d}{||\mathbf{q}_j^d||_2} - \frac{\mathbf{q}_j^p}{||\mathbf{q}_j^p||_2} ||_2, \tag{9}$$

where $\mathbf{q}_j^d$ and $\mathbf{q}_j^p$ are the jth attention maps from S_d and the trained DFEM in the vectorized forms, respectively.

As mentioned previously, each FER database involves certain types of disturbing factors. In the ADFL, the SA layer estimates the importance weights corresponding to different disturbing factors based on $\mathbf{f}^d$, while the feature fusion layer outputs $\mathbf{f}^m$ based on these importance weights. Therefore, $\mathbf{f}^m$ incorporates the prior disturbance information that considers the characteristics of the FER database. For instance, the prior pose information does not greatly contribute to the extraction of $\mathbf{f}^m$ for the in-the-lab FER database (since the importance weight corresponding to pose is low in this case), while it is encoded in $\mathbf{f}^m$ for the in-the-wild FER database. By performing adversarial transfer learning, the distributions of $\mathbf{f}^m$ and $\mathbf{f}^d$ are as similar as possible. Hence, we are able to effectively extract adaptive disturbance-specific features, which approximate a linear combination of disturbing factor features from the trained DFEM, by exploiting the knowledge of the FER database.

In the SA layer, the importance weight reflects the influence of a disturbing factor. To explain this, we take a simple example for illustration. Assume that we have an FER training database only involving identity variations and that α_1 corresponds to the importance weight of the identity. In other words, all the images in this FER database are captured with the same gender, age, race, illumination, and pose. Thus, the first disturbing factor features (i.e., $\mathbf{f}_1^p$ corresponding to identity) of different images in the FER database significantly vary, while the others (i.e., $\{\mathbf{f}_2^p, \ldots, \mathbf{f}_M^p\}$) show small variations. By minimizing the differences between $\mathbf{f}^d$ and $\mathbf{f}^m$, α_1 and α_j ($j \in \{2, \cdots, M\}$)) are assigned large and small values, respectively. Accordingly, the joint loss function [see Eq. (14)] can be gradually optimized. Otherwise, (i.e., α_1 is small while α_j ($j \in \{2, \ldots, M\}$)) is large), the weighted disturbing factor features [see Eq. (6)] are similar for all the images. In such a case, the disturbance subnetwork fails to extract effective information, and thus disturbance disentanglement cannot be properly performed. Therefore, the value of α_1 reflects the influence of identity.

Note that the multi-level attention mechanism is not used in S_d. This is because the salient regions for identifying multiple disturbing factors are different. For example, illumination estimation mainly relies on the global facial region, while pose classification focuses on local regions around salient facial landmarks. In other words, it is not appropriate to directly concatenate low-level spatial features and high-level semantic features to extract disturbance-specific features in S_d.

3.3.3 Attention Block

Inspired by Liu et al. (2019), we develop an attention block for both S_d and S_e. The network architecture of the attention block is given in Fig. 4.

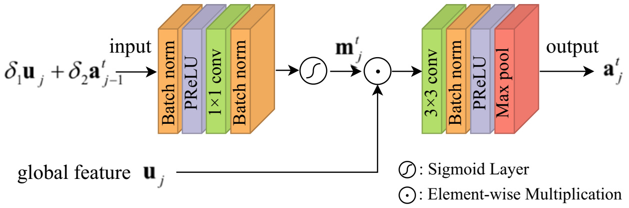


Fig. 4 The network architecture of the attention block

The first attention block in S_e or S_d takes the feature $\mathbf{u}_1$ from the first convolution block in S_g as the input. For the subsequent attention block at the jth layer, the element-wise weighted addition between the global feature $\mathbf{u}_j$ from S_g and the task-specific feature $\mathbf{a}^t_{j-1}$ ($t \in \{e, d\}$) from the previous layer in S_t ($t \in \{e, d\}$) is taken as the input, as shown in Fig. 4. Then, the attention mask $\mathbf{m}^t_j$ ($t \in \{e, d\}$) generated from the jth layer in S_t ($t \in \{e, d\}$) is expressed as

$$\mathbf{m}^t_j = \begin{cases} g(\mathbf{u}_j), & j = 1, \\ g(\delta_1\mathbf{u}_j + \delta_2\mathbf{a}^t_{j-1}), & j \geq 2, \end{cases} \tag{10}$$

where δ_1 and δ_2 are the learnable parameters that, respectively, determine the importance of the global feature $\mathbf{u}_j$ and the task-specific feature $\mathbf{a}^t_{j-1}$; $g(\cdot)$ denotes the aggregation of a batch normalization (BN) layer, a parametric ReLU (PReLU) 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 map $\mathbf{a}^t_j$ of the jth attention block for S_t ($t \in \{e, d\}$) is given as

$$\mathbf{a}^t_j = h(\mathbf{m}^t_j \odot \mathbf{u}_j), \tag{11}$$

where '$\odot$' denotes the element-wise multiplication; $h(\cdot)$ denotes a convolutional layer with a 3×3 kernel that matches the channels between the attention mask from the $(j-1)$th layer in S_t ($t \in \{e, d\}$) and the global shared feature in the jth layer in S_g, followed by a BN layer, a PReLU layer, and a max pooling layer to match the sizes of the feature maps between the above two features.

It is worth noting that our attention block outputs a 3D attention mask, where each attention map in the mask captures salient regions for a feature channel in S_g. This is different from the traditional attention block (Xie et al., 2019a), which applies the same 2D mask to each feature channel. Therefore, the attention block used in our paper takes into account the differences between feature maps and thus can generate more accurate attention weights.

3.3.4 Mutual Information Neural Estimator (MINE)

To perform explicit disentanglement between the disturbance-specific feature $\mathbf{f}^d$ and the expression-specific feature $\mathbf{f}^e$, the correlation between the two features should be minimized. Generally, the Kullback–Leibler (K–L) divergence $\mathcal{D}_{KL}(\mathbb{P}_{F^d}||\mathbb{P}_{F^e})$ can be used to minimize the discrepancy between two feature distributions. Here, F^d and F^e denote the random variables of $\mathbf{f}^d$ and $\mathbf{f}^e$, respectively. $\mathbb{P}_{F^d}$ and $\mathbb{P}_{F^e}$ represent the marginal probability distributions of F^d and

F^e, respectively. However, we cannot guarantee that the features with dissimilar distributions are uncorrelated.

Inspired by Belghazi et al. (2018), we leverage mutual information to measure the correlation between $\mathbf{f}^d$ and $\mathbf{f}^e$ (note that if two variables are independent of each other, their mutual information is zero). Specifically, we employ a mutual information neural estimator (MINE) (Belghazi et al., 2018) to estimate the mutual information between $\mathbf{f}^d$ and $\mathbf{f}^e$, leading to explicit disentanglement. Based on the K–L divergence and the Donsker–Varadhan representation (Donsker & Varadhan, 1983), the mutual information can be estimated by the MINE as

$$\begin{aligned} I(F^d; F^e) &= \mathcal{D}_{KL}(\mathbb{P}_{F^d F^e} \| \mathbb{P}_{F^d} \otimes \mathbb{P}_{F^e}) \\ &\geq \mathbb{E}_{\mathbb{P}_{F^d F^e}}[T_\theta(\mathbf{f}^d, \mathbf{f}^e)] \\ &\quad - \log(\mathbb{E}_{\mathbb{P}_{F^d} \otimes \mathbb{P}_{F^e}}[e^{T_\theta(\mathbf{f}^d, \mathbf{f}^e)}]), \end{aligned} \tag{12}$$

where '$\otimes$' is the product function; $\mathbb{P}_{F^d F^e}$ represents the joint probability distribution of (F^d, F^e); and T_θ is a neural network with parameters θ (the detailed architecture of T_θ is described in Table 1a).

Given n mini-batch samples $\{\mathbf{f}_i^d, \mathbf{f}_i^e\}_{i=1}^n$ from the joint distribution and n samples $\{\tilde{\mathbf{f}}_i^e\}_{i=1}^n$ from the marginal distribution of F^e (which can be estimated by shuffling the samples from the joint distribution along the batch axis), the mutual information loss $\mathcal{L}_{MI}$ is approximated as

$$\begin{aligned} \mathcal{L}_{MI} &= I(F^d; F^e) \\ &\approx \frac{1}{n} \sum_{i=1}^{n} T_\theta(\mathbf{f}_i^d, \mathbf{f}_i^e) - \log\left(\frac{1}{n} \sum_{i=1}^{n} e^{T_\theta(\mathbf{f}_i^d, \tilde{\mathbf{f}}_i^e)}\right). \end{aligned} \tag{13}$$

The correlation between $\mathbf{f}^d$ and $\mathbf{f}^e$ is minimized by optimizing the mutual information loss $\mathcal{L}_{MI}$. Therefore, we are able to disentangle the disturbance in an explicit way.

3.3.5 Joint Loss Function

The joint loss function of the ADDM is defined as

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{AD} + \lambda_2 \mathcal{L}_{AT} + \lambda_3 \mathcal{L}_{MI}, \tag{14}$$

where λ_1, λ_2, and λ_3 denote the balanced parameters of the adversarial loss, the attention transfer loss, and the mutual information loss, respectively.

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

3.4 Discussions

A number of CNN-based FER methods (Mollahosseini et al. 2016; Yu & Zhang, 2015) suffer from the problem that the final expression features contain the disturbance because of limited training data. Some disturbance-disentangled-based FER methods (Zhang et al., 2018b; Meng et al., 2017) may not accurately recognize expressions in the disturbance-unlabeled FER database.

Different from traditional FER methods, the ADDL method successfully leverages the available disturbance label information from the large-scale face database to perform adversarial transfer learning on the disturbance-unlabeled FER database. In particular, by designing a disturbance subnetwork and minimizing the mutual information, the disturbance can be effectively and explicitly disentangled from the features used for expression recognition. Such a manner significantly improves the discriminability of expression-specific features. Therefore, the problems due to limited training data and the lack of disturbance labels can be greatly alleviated. Moreover, the ADFL is developed to facilitate the extraction of disturbance-specific features by fully exploiting the different influences of disturbing factors in the FER database.

4 Experiments

In this section, extensive experiments are conducted to show the superiority of our proposed method. First, we introduce several public FER databases and the implementation details in Sects. 4.1 and 4.2, respectively. Then, we conduct ablation studies to evaluate each component of our proposed method in Sect. 4.3. Next, we compare our method with several state-of-the-art FER methods in Sect. 4.4. Finally, we present the computational complexity of our method and apply our method to valence and arousal estimation in Sects. 4.5 and 4.6, respectively.

4.1 Databases

To validate the effectiveness of the proposed method, we evaluate the performance on three in-the-lab FER databases [CK+ (Lucey et al., 2010), MMI (Valstar & Pantic, 2010), and Oulu-CASIA (Zhao et al., 2011)) and four in-the-wild databases (RAF-DB (Li et al., 2017), SFEW (Dhall et al., 2011), Aff-Wild2 (Kollias & Zafeiriou, 2018), and AffectNet (Mollahosseini et al. 2017)].

CK+: The Extended Cohn-Kanade (CK+) database is a commonly used laboratory-controlled database for evaluating the FER performance. It contains 327 video sequences annotated with expression labels, including six basic expressions (i.e., angry, happy, surprise, sad, disgust, and fear) and one nonbasic expression (i.e., contempt). Each sequence shows a shift from a neutral expression to a peak expression. We choose the last three expressional frames from each sequence to con-

Table 1 The detailed architecture of the MINE and each subnetwork in the ADDM

MINE	Output dimensionality
(a) The architecture of the MINE	
Concatenate[1]	256
FC(64), Leaky ReLU	64
FC(1), Leaky ReLU	1

The global shared subnetwork S_g	Output dimensionality
(b) The architecture of the global shared subnetwork in the ADDM	
Conv(64, 7, 2), BN, ReLU, Max pool(3,2)	64 × 56 × 56
Basic block(64), Basic block(64)	64 × 56 × 56
Basic block(128), Basic block(128)	128 × 28 × 28
Basic block(256), Basic block(256)	256 × 14 × 14
Basic block(512), Basic block(512)	512 × 7 × 7

The expression subnetwork S_e	Output dimensionality
(c) The architecture of the expression subnetwork in the ADDM	
Attention block(64)	64 × 56 × 56
Max pool(10,7)	64 × 7 × 7
Attention block(128)	128 × 28 × 28
Max pool(8,3)	128 × 7 × 7
Attention block(256)	256 × 14 × 14
Max pool(2,2)	256 × 7 × 7
Attention block(512)	512 × 7 × 7
Attention block(512)	512 × 7 × 7
Concatenate[2]	1472 × 7 × 7
Avg pool(3)	1472 × 2 × 2
Flatten	5888
Dropout, FC(128), PReLU	128
Dropout, FC(K), PReLU	K

The disturbance subnetwork S_d	Output dimensionality
(d) The architecture of the disturbance subnetwork in the ADDM	
Attention block(64)	64 × 56 × 56
Attention block(128)	128 × 28 × 28
Attention block(256)	256 × 14 × 14
Attention block(512)	512 × 7 × 7
Attention block(512)	512 × 7 × 7
Avg pool(6)	512 × 1 × 1
Flatten	512
Dropout, FC(128), PReLU	128
FC(6), Sigmoid	6

Conv(n, m, s) denotes the convolutional layer with the number of output feature maps n, the kernel size $m \times m$ and the stride s; Basic block (n) and Attention block (n), respectively, denote the basic block and the attention block with the number of output feature maps n; Max pool (m,s) denotes the max pooling layer with $m \times m$ filters and s strides; Concatenate[1] denotes the concatenation of $\mathbf{f}^d$ and $\mathbf{f}^e$ in the MINE; Concatenate[2] denotes the concatenation of all the outputs of attention blocks in S_e; Avg pool(m) denotes the average pooling layer with $m \times m$ filters; FC(n) refers to the fully-connected layer with the output features of n dimensions; The value of K refers to the number of classes; BN denotes a batch normalization layer; PReLU denotes a parametric ReLU layer struct the training set and the test set, which contain 981 images in total.

MMI: The MMI database is composed of 30 subjects, for which 205 image sequences captured in the frontal view are labeled with six basic facial expressions. Similar to the CK+ database, we select the three peak expressional frames in each sequence to compose the training set and the test set (consisting of 615 images in total).

Oulu-CASIA: The Oulu-CASIA database contains videos of 80 subjects. Each subject contains six basic expressions, where each expression corresponds to a video sequence. The videos are collected with two imaging systems (i.e., near-infrared and visible light) under three different illumination conditions. As done in Yang et al. (2018a), the last three frames in each sequence captured with visible light and strong illumination are used in our experiments, resulting in a total of 1,440 images.

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 labels, the images in RAF-DB 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, which covers unconstrained facial expressions, varied head poses, large age range, varied focus, different resolutions of faces, and real-world illumination. It provides 958 images for training and 436 images for testing. Each image is labeled with one of six basic expressions or the neutral expression.

Aff-Wild2: The Aff-Wild2 database is extended from the Aff-Wild database (Kollias et al., 2019), which consists of 558 YouTube videos with 2,786,201 frames. The videos involve large variations in age, race, pose, illumination, and so on. In this paper, we use the preprocessed version provided by Zhang et al. (2020c) in the ABAW 2020 competition (Kollias et al., 2020b), which contains 904,825 images for training and 322,080 validation images for testing. All the images are annotated with seven expression categories, as in RAF-DB and SFEW.

AffectNet: The AffectNet database is a large-scale database of facial emotions in the wild. It contains 450,000 facial images from the Internet with both categorical (including seven expressions) and valence-arousal annotations. For FER, we select 283,901 images for training and 3500 validation images for testing, as done in Zeng et al. (2018), Wang et al. (2019), Farzaneh and Qi (2021). For valence and arousal estimation, we use all the images with valence-arousal annotations, resulting in 320,739 images for training and 4,500 images for testing.

For in-the-lab databases, we employ the popular tenfold cross-validation protocol for evaluation, as done in Meng et al. (2017), Yang et al. (2018a), Zhao et al. (2016), Ding et al. (2017). For in-the-wild databases, we follow the default evaluation protocols provided by the databases.

4.2 Implementation Details

In this paper, we use ResNet-18 pretrained on the MS-Celeb-1M database as the backbone (Wang et al., 2020b). The dimensionalities of $\{\mathbf{f}_j^p\}_{j=1}^M$, $\mathbf{f}^d$, and $\mathbf{f}^e$ are 128. Table 1 illustrates the detailed architecture of the MINE and each subnetwork in the ADDM, where the output dimensionality of each layer is also given.

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 256 × 256. During training, the facial images are randomly cropped to the size of 224 × 224, and the cropped images are further processed by using a horizontal flip. For the Aff-Wild2 and AffectNet databases, the oversampling strategy is used, as done in Wang et al. (2020b).

Since five blocks are used in ResNet-18, L is set to five in Eqs. (3) and (9). The values of λ_1, λ_2, and λ_3 in Eq. (14) are empirically set to 1.0, 0.10, and 0.0010, respectively. We train the networks using the Adam algorithm (Kingma & Ba, 2014) with a 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 our models are trained on a single NVIDIA GTX 1080Ti GPU using PyTorch for 40 epochs, with a batch size of 16 for RAF-DB and AffectNet and 8 for the other FER databases (except for Aff-Wild2). For Aff-Wild2, our model is trained on two NVIDIA GTX 1080Ti GPUs for 40 epochs with a batch size of 64. For computational complexity, we evaluate the inference time and speed of our method by using a single NVIDIA GTX 1080Ti GPU.

The DFEM is trained on both the Multi-PIE face database (Gross et al., 2010) and the RAF-DB database, which provide the labels of multiple disturbing factors. Note that large-scale facial attribute databases (such as CelebA (Liu et al., 2015)) are not used for training. This is because they do not have labels of illumination and pose (which are not facial attributes). Moreover, CelebA only contains binary facial attributes (with and without), and thus, it cannot comprehensively describe the variations of each attribute. In contrast, Multi-PIE has labels of identity (337 individuals), pose (15 viewpoints), and illumination (19 lighting conditions), while RAF-DB gives those of gender (3 classes), age (5 ranges), and race (3 classes). Therefore, Multi-PIE and RAF-DB are more suitable to train the DFEM. During the training of the DFEM, missing labels of some disturbing factors are ignored during back-propagation.

For the Aff-Wild2 database, we use the weighted average of accuracy (33%) and F1 score (67%) as the evaluation metric, as done in the ABAW 2020 competition (Kollias et al.,

Table 2 Details of the three baseline methods, six DDL variants, and four ADDL variants

Methods	S_g	S_e		S_d						ADFL	MI	DFEM
		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	✓	–	✓	✓	✓	✓	✓	✓	✓	–	–	✓
ADDL_ADFL	✓	–	✓	✓	✓	✓	✓	✓	✓	✓	–	✓
ADDL_MI	✓	–	✓	✓	✓	✓	✓	✓	✓	–	✓	✓
ADDL_MI-DFEM	✓	–	✓	–	–	–	–	–	–	–	✓	–
ADDL	✓	–	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

"w/o" and "Multi" represent training without and with the multi-level attention mechanism, respectively; "Gen", "Id" and "Ill" represent gender, identity, and illumination, respectively; and "MI" represents explicit disentanglement using mutual information

2020b). For the other databases, we adopt the test accuracy as the evaluation metric.

4.3 Ablation Studies

To show the superiority of the proposed method, we perform extensive ablation studies to evaluate the influence of different components on the performance. In this subsection, we use one in-the-lab database (MMI) and one in-the-wild database (RAF-DB) for evaluation.

Specifically, we evaluate the performance of three baseline methods, six DDL variants, and four ADDL variants. DDL refers to our original method (Ruan et al., 2020) that does not involve the ADFL and the MINE, while the ADDL method is developed in this paper.

These methods are described as follows: (1) The baseline method (denoted Baseline) that uses only S_g followed by two FC layers to predict the expression of the input image. (2) The baseline method with attention blocks (denoted Baseline_at) that simultaneously uses S_g and S_e, but does not use the multi-level attention mechanism in S_e. (3) The baseline method with attention blocks (denoted Baseline_mat) that simultaneously uses S_g and S_e, and employs the multi-level attention mechanism in S_e. (4) The method (denoted DDL_g) that simultaneously uses S_g, S_e, and S_d, where S_d is trained based on the gender features extracted by the DFEM. (5) The method (denoted DDL_ga) that is similar to DDL_g, but where S_d is trained based on both the gender and age features extracted by the DFEM. (6) The method (denoted DDL_gar) that is similar to DDL_g, but where S_d is trained based on the gender, age, and race features extracted by the DFEM. (7) The method (denoted DDL_gar&id) that is similar to DDL_g, but where S_d is trained based on the gender, age, race, and identity features extracted by the DFEM. (8) The method (denoted DDL_gar&id&il) that is similar to DDL_g, but where S_d is trained based on the gender, age, race, identity, and illumination features extracted by the DFEM. (9) The method (denoted DDL_gar&id&il&p) that is similar to DDL_g, but where S_d is trained based on the gender, age, race, identity, illumination, and pose features extracted by the DFEM. (10) The ADDL method (denoted ADDL_ADFL) that only uses the ADFL. (11) The ADDL method (denoted ADDL_MI) that employs only the MINE to perform explicit disentanglement between $\mathbf{f}^d$ and $\mathbf{f}^e$, where $\mathbf{f}^d$ is learned using the DFEM as for the DDL_gar&id&il&p. (12) The ADDL method (denoted ADDL_MI-DFEM) that employs the MINE but without using the DFEM. (13) The ADDL method that simultaneously uses the ADFL and the MINE.

The details of these methods are summarized in Table 2. For a fair comparison, the pretrained ResNet-18 is employed for all the methods. Table 3 reports the recognition accuracy obtained by these methods on the MMI and RAF-DB databases.

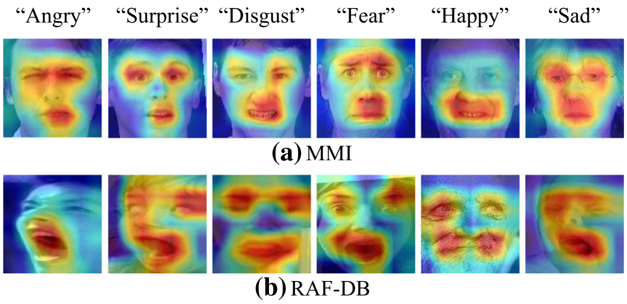
4.3.1 Posed versus Naturalistic Facial Expressions

Generally, in-the-lab FER databases contain posed facial expressions, while in-the-wild databases are comprised of naturalistic facial expressions. Posed expressions usually have slow and jerky onsets, where facial actions typically do not show peaks simultaneously (Motley & Camden, 1988).

Table 3 The recognition accuracy (%) obtained by the three baseline methods, six DDL variants, and four ADDL variants on the MMI and RAF-DB databases

Methods	MMI	RAF-DB
Baseline	79.16	86.93
Baseline_at	81.22	87.45
Baseline_mat	81.47	87.61
DDL_g	82.88	87.97
DDL_ga	83.29	88.01
DDL_gar	83.70	88.07
DDL_gar&id	83.74	88.10
DDL_gar&id&il	84.25	88.14
DDL_gar&id&il&p	83.56	88.17
ADDL_ADFL	85.40	89.08
ADDL_MI	85.56	88.82
ADDL_MI-DFEM	83.13	88.01
ADDL	**86.13**	**89.34**

The best results are boldfaced



(a) MMI

(b) RAF-DB

Fig. 5 Visualization of attentive feature maps on the **a** MMI and **b** RAF-DB databases

In contrast, naturalistic expressions tend to exhibit fast and smooth onsets, where distinct facial movements reach peaks in a short duration. According to Table 3, compared with the baseline, the accuracy gains obtained by the ADDL method are 6.97% and 2.41% on the MMI and RAF-DB databases, respectively. This shows the importance of disentangling disturbance for both the posed and naturalistic FER, which enables the extraction of effective expression-specific features. Note that the recognition accuracy obtained by our method on MMI is lower than that on RAF-DB. This can be ascribed to the limited training set (note that there are 615 images in MMI), thereby increasing the difficulty of learning a robust FER model.

4.3.2 Influence of the Attention Block and Multi-level Attention Mechanism

As illustrated in Table 3, Baseline_at achieves better recognition performance than the Baseline method on both MMI and RAF-DB. Specifically, compared with Baseline, Baseline_at achieves 2.06% and 0.52% gains in terms of recognition accuracy on the MMI and RAF-DB databases, respectively. The above results show the effectiveness of the attention block.

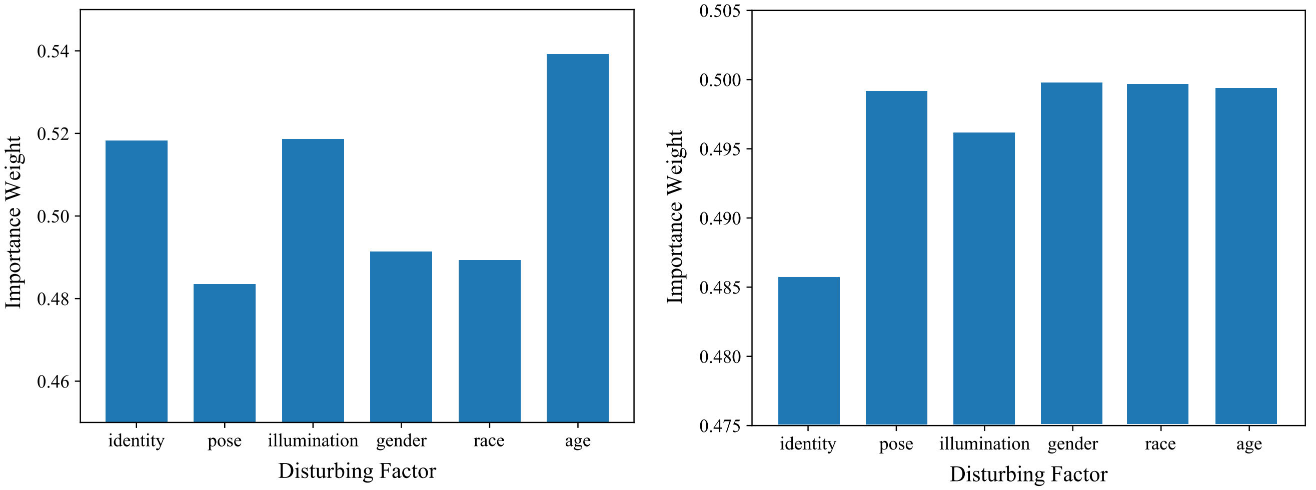
Baseline_mat obtains higher recognition accuracy than Baseline_at. Specifically, in comparison with Baseline_at, Baseline_mat gets 0.25% improvements in terms of recognition accuracy on MMI. For RAF-DB, its accuracy is further improved by 0.16%. This verifies the effectiveness of the multi-level attention mechanism.

To further 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 Fig. 5. Specifically, the combined feature maps [see Eq. (3)] before the FC layer are first added along the channel dimension, which generates an attentive feature map with a size of 7×7. 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 result.

As given in Fig. 5, the warm-toned parts of an image correspond to the regions with large values in the attentive feature map, while the cold-toned parts correspond to the regions with small values in the attentive feature map. We can observe that the attentive feature map is able to focus on the salient facial regions (especially the regions around the eyes and mouth) that are critical for FER. In particular, for the images in RAF-DB, the corresponding attentive feature maps tend to focus on larger facial patches than those in MMI. This is because the images in RAF-DB involve large pose variations and low quality. A larger facial patch is beneficial to extract more discriminative features for FER on the in-the-wild database.

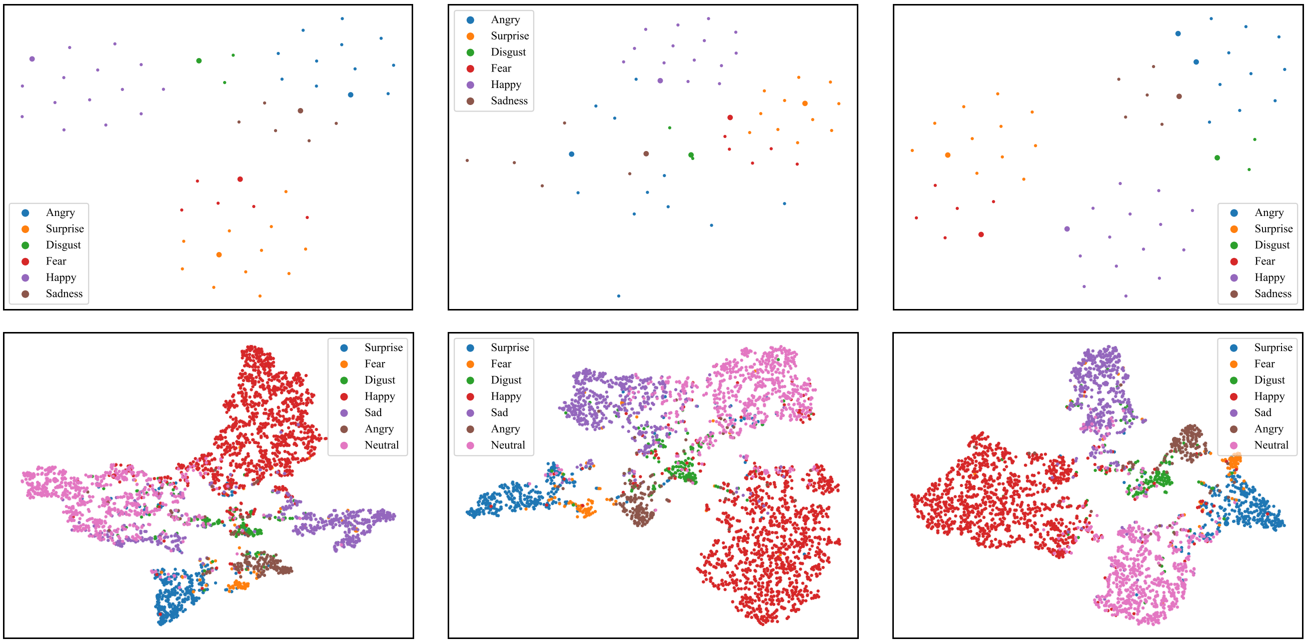
4.3.3 Influence of the Different Disturbing Factors

As shown in Table 3, all the DDL variants consistently perform better than Baseline_mat, which demonstrates the importance of the disturbance subnetwork S_d. For the RAF-DB database, the recognition accuracy obtained by DDL tends to be higher when more disturbing factors are considered. DDL achieves the best performance when all the disturbing factors are employed for disturbance-specific feature learning in S_d. This is because the images in RAF-DB contain severe variations caused by multiple disturbing factors. Disentangling these disturbing factors from facial expression images benefits the extraction of effective expression-specific features. However, for the MMI database, DDL obtains the best accuracy when all the disturbing factors except for the pose are considered. This is because the images in MMI do not involve pose variations (the images are all frontal). Therefore, it is critical to prop-



(a) MMI **(b)** RAF-DB

Fig. 6 Visualization of the importance weights (corresponding to various disturbing factors) learned by the proposed ADDL in the training sets of the **a** MMI and **b** RAF-DB databases



(a) Baseline **(b)** Baseline_mat **(c)** ADDL

Fig. 7 Feature visualization using t-SNE. The features are extracted by using two baseline methods and the proposed ADDL method. The first row shows the feature visualization on the MMI database, and the second row shows the feature visualization on the RAF-DB database. **a** Feature visualization on the model trained by Baseline. **b** Feature visualization on the model trained by Baseline_mat. **c** Feature visualization on the model trained by the ADDL method

erly choose the disturbing factors by taking into account the characteristics of the FER database.

4.3.4 Influence of the ADFL and MINE

From Table 3, we can make the following observations. First, the ADDL_ADFL method achieves better recognition performance than all the DDL variants. Compared with the DDL_gar&id&il&p method, which does not consider the different influences of disturbing factors (i.e., the importance weights corresponding to all the disturbing factors are the same), the ADDL_ADFL method obtains 1.84% and 0.91% improvements in terms of recognition accuracy on MMI and RAF-DB, respectively. This indicates that adopting the SA layer is effective in learning the importance weights, which can be further beneficial to the extraction of disturbance-specific features in S_d.

Table 4 The NMI values obtained by different methods

Methods	MMI	RAF-DB
Baseline	0.610	0.655
Baseline_at	0.628	0.671
Baseline_mat	0.662	0.671
DDL_g	0.671	0.676
DDL_ga	0.675	0.678
DDL_gar	0.676	0.680
DDL_gar&id	0.682	0.682
DDL_gar&id&il	0.688	0.682
DDL_gar&id&il&p	0.670	0.685
ADDL_ADFL	0.693	0.699
ADDL_MI	0.700	0.694
ADDL_MI-DFEM	0.672	0.679
ADDL	**0.711**	**0.709**

For NMI values, the higher is better. The best results are boldfaced

Table 5 Ablation studies for the influence of the different DFEM models

Databases	MMI	RAF-DB
Multi-PIE	85.00	88.89
RAF-DB	85.71	88.69
Multi-PIE & RAF-DB	**86.13**	**89.34**

The best recognition accuracies (%) are boldfaced

Second, to demonstrate the importance of explicit disentanglement, we jointly train the MINE and the ADDM in the ADDL_MI method. The ADDL_MI method also obtains higher accuracy than all the DDL variants. Therefore, minimizing the mutual information is advantageous to explicitly disentangle disturbance-specific features from expression-specific features and has a positive influence on the final performance.

Third, the ADDL method achieves the best accuracy on both in-the-lab and in-the-wild databases when both the ADFL and MINE are jointly adopted. Specifically, the proposed ADDL method outperforms the DDL_gar&id&il&p method by 2.57% and 1.17% on MMI and RAF-DB, respectively. In summary, the developed ADFL and MINE are effective to improve the FER performance.

To illustrate that the importance weights from the AFDL can reflect the different influences of disturbing factors in the FER training database, we visualize the importance weights learned by the ADDL method in the training sets of MMI and RAF-DB, as shown in Fig. 6. In Fig. 6a, the importance weight corresponding to the pose is smaller than those corresponding to the other disturbing factors in MMI. This is because the images from MMI do not contain severe pose variations. In Fig. 6b, the weights corresponding to gender, race, age, and pose are similar and higher than the weight corresponding to identity in RAF-DB. This indicates that RAF-DB suffers from more disturbing factors than MMI. Therefore, the above results validate that the proposed AFDL can adaptively estimate the importance weights corresponding to different disturbing factors according to the characteristics of the FER database.

To demonstrate that our proposed method is able to extract discriminative features for expression recognition, we further use t-SNE (Maaten & Hinton, 2008) to visualize the features in the 2D space. Figure 7 shows the feature visualization obtained by the Baseline, Baseline_mat, and ADDL methods on the MMI and RAF-DB databases.

From Fig. 7, we can see that the proposed ADDL method can effectively reduce intra-class variances and inter-class similarities compared with Baseline and Baseline_mat. Baseline_mat achieves better inter-class separation and intra-class compactness than Baseline, which verifies the superiority of the multi-level attention mechanism and attention blocks. As shown in the second row of Fig. 7, due to the great challenges of the RAF-DB database, the features from different classes severely overlap for the Baseline method. In contrast, for the proposed ADDL method, the features from the same class are more closely clustered, while the inter-class distances are enlarged (especially for surprise, sad, neutral, and disgust expressions). Therefore, our method is capable of effectively disentangling the disturbance, even when some challenging variations occur in facial expression images.

Finally, we adopt the Normalized Mutual Information (NMI) value to quantitatively measure the quality of classification results obtained by different methods, as shown in Table 4. We can observe that the ADDL method gives the highest NMI value among all the competing methods. Moreover, both ADDL_ADFL and ADDL_MI obtain higher NMI values than the three baselines and six DDL variants. This further demonstrates the effectiveness of ADFL and MINE for reducing intra-class differences and inter-class similarities.

4.3.5 Influence of the DFEM

We evaluate the influence of the DFEM on the final performance. We compare the ADDL_MI-DFEM with the ADDL and ADDL_ADFL methods.

As shown in Table 3, we can see that the recognition accuracy obtained by the ADDL_MI-DFEM method significantly drops on both in-the-lab and in-the-wild databases compared with that of the ADDL method. When the DFEM is not adopted, ADDL is only optimized by the CE loss and mutual information loss. In this way, the disturbance-specific features are learned in an unsupervised way without using any prior knowledge about disturbing factors.

Table 6 Ablation studies for the influence of backbones pretrained on different databases

Pretrained	Databases	MMI	RAF-DB
–	–	70.76	85.30
✓	ImageNet	84.03	87.58
✓	AffectNet	85.30	88.46
✓	MS-Celeb-1M	**86.13**	**89.34**

The best recognition accuracies (%) are boldfaced

Table 7 Ablation studies for the influence of the balanced parameters λ_1, λ_2, and λ_3 on the MMI and RAF-DB databases

λ_1	MMI	RAF-DB
(a) Influence of λ_1		
0.0	83.39	88.27
0.5	84.72	88.92
1.0	**86.13**	**89.34**
1.5	85.50	88.62
2.0	83.45	88.30
λ_2	MMI	RAF-DB
(b) Influence of λ_2		
0.00	83.28	88.04
0.05	85.45	88.85
0.10	**86.13**	**89.34**
0.15	85.08	89.11
0.20	84.74	88.98
λ_3	MMI	RAF-DB
(c) Influence of λ_3		
0.0000	85.40	89.08
0.0001	85.77	89.05
0.0010	**86.13**	**89.34**
0.0100	85.67	89.08
0.1000	84.91	88.82

The best recognition accuracies (%) are boldfaced

Thus, the disturbance subnetwork cannot effectively capture disturbance-related information, degrading the disentanglement performance of ADDL.

Compared with the ADDL_MI-DFEM, the ADDL_ADFL method also has better performance since it is able to extract more discriminative disturbance-specific features by leveraging prior information based on the trained DFEM. Therefore, the DFEM plays a critical role in the disturbance disentanglement to improve the accuracy of the ADDL method.

4.3.6 Influence of the Different DFEM Models

We evaluate the performance of our method with the different DFEM models trained based on three face databases (Multi-PIE, RAF-DB, and Multi-PIE & RAF-DB), as shown in Table 5.

We can see that when the DFEM is trained based on Multi-PIE (including the labels of identity, illumination, and pose), ADDL only achieves 85.00% and 88.89% on MMI and RAF-DB, respectively. When the DFEM is trained based on RAF-DB (including the labels of gender, race, and age), ADDL obtains 85.71% and 88.69% on MMI and RAF-DB, respectively. However, when both RAF-DB and Multi-PIE are used to train the DFEM, the performance of the ADDL method is greatly improved. This further shows the importance of considering different types of disturbing factors for disturbance disentanglement.

4.3.7 Influence of Backbones Pretrained on the Different Databases

We investigate the influence of backbones pretrained on the different databases (including ImageNet, AffectNet, and MS-Celeb-1M) on the final performance, as shown in Table 6. The performance obtained by our method without pretraining the backbone is also evaluated.

Our method with the pretrained backbone achieves much better FER performance than that without pretraining the backbone. Moreover, our method with the backbone pretrained on MS-Celeb-1M gives better performance than those pretrained on other large-scale databases. This is because MS-Celeb-1M (including 10 M images) contains many more facial images than AffectNet (including 283 K images), which facilitates the backbone network to extract more effective global features for FER. Although there are a great number of images in ImageNet, most samples are natural images rather than facial images. Therefore, our method with the backbone pretrained on ImageNet gives the worst performance among the three pretrained backbones.

4.3.8 Influence of Balanced Parameters

We study the influence of three balanced parameters (i.e., λ_1, λ_2, and λ_3) in the joint loss [Eq. (14)], as shown in Table 7.

Specifically, we first fix $\lambda_2 = 0.10$ and $\lambda_3 = 0.0010$, and set the values of λ_1 from 0.0 to 2.0. The results are shown in Table 7a. When $\lambda_1 = 0.0$, adversarial training is not adopted, and thus, the disturbance-specific features cannot be effectively learned, leading to a performance decrease. When $\lambda_1 = 1.0$, the proposed method obtains the highest accuracy. Then, we fix $\lambda_1 = 1.0$ and $\lambda_3 = 0.0010$, and set the values of λ_2 from 0.00 to 0.20. The results are shown in Table 7b. The proposed method achieves the top performance when $\lambda_2 = 0.10$. Note that when $\lambda_2 = 0.00$ (attention transfer is not used in this case), the proposed method achieves worse accuracy than that without using adversarial training on both MMI and RAF-DB. Hence, it is important to bridge the gap

Table 8 Comparisons of all the competing methods on in-the-lab databases (CK+, MMI, and Oulu-CASIA)

Methods	Accuracy (%)		
	CK+	MMI	Oulu-CASIA
LBP-TOP (Zhao & Pietikainen, 2007)	88.99‡	59.51	68.13
DTAGN* (Jung et al., 2015)	97.25‡	70.20	81.46
PPDN (Zhao et al., 2016)	97.30†	–	72.40
IACNN (Meng et al., 2017)	95.37‡	71.55	–
PHRNN-MSCNN* (Zhang et al., 2017)	98.50‡	81.18	86.25
FN2EN (Ding et al., 2017)	98.60†	–	87.71
DLP-CNN (Li & Deng, 2018)	95.78†	78.46	-
DeRL (Yang et al., 2018a)	97.37‡	73.23	88.00
IPA2LT (Zeng et al., 2018)	92.45‡	65.61	61.49
DAM-CNN (Xie et al., 2019a)	95.88†	–	–
L2-sparseness (Xie et al., 2019b)	97.59‡	78.54	82.92
FMPN (Chen et al., 2019)	98.06	82.74	–
TDGAN (Xie et al., 2020)	$97.53 \pm 2.03^{\ddagger}$	–	–
DDL (Ruan et al., 2020)	99.16‡	83.67	88.26
ADDL (proposed)	**99.64**‡	**86.13**	**89.44**

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

between the DFEM and the disturbance subnetwork at the lower layers. Finally, Table 7c illustrates the results obtained by our method by fixing $\lambda_1 = 1.0$ and $\lambda_2 = 0.10$ and varying the values of λ_3 from 0.0000 to 0.1000. We can observe that the proposed method obtains the best accuracy when $\lambda_3 = 0.0010$. In this paper, we use $\lambda_1 = 1.0$, $\lambda_2 = 0.10$, and $\lambda_3 = 0.0010$ for all the experiments.

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

In this subsection, we compare our proposed method with several state-of-the-art FER methods.

For in-the-lab databases, we compare the proposed ADDL with fourteen representative FER methods, including LBP-TOP (Zhao & Pietikainen, 2007), PPDN (Zhao et al., 2016), FN2EN (Ding et al., 2017), IACNN (Meng et al., 2017), DLP-CNN (Li & Deng, 2018), DTAGN (Jung et al., 2015), DeRL (Yang et al., 2018a), IPA2LT (Zeng et al., 2018), DAM-CNN (Xie et al., 2019a), PHRNN-MSCNN (Zhang et al., 2017), L2-sparseness (Xie et al., 2019b), FMPN (Chen et al., 2019), TDGAN (Xie et al., 2020), and our previous DDL (Ruan et al., 2020). For in-the-wild databases, we also compare our proposed ADDL with several representative FER methods, including gACNN (Li et al., 2018), IPA2LT (Zeng et al., 2018), SPDNet (Acharya et al., 2018), IPFR (Wang et al., 2019), RAN (Wang et al., 2020c), SCN (Wang et al., 2020b), FMPN (Chen et al., 2019), CPG (Hung et al., 2019b), PAENet (Hung et al., 2019a), PSR (Vo et al., 2020), DACL (Farzaneh & Qi, 2021), EfficientNet-B0 (Savchenko, 2021), CNN (Anas et al., 2020), NISL (Deng et al., 2020), LLAM (Wang et al., 2020a), ICT-VIPL (Zhang et al., 2020c), DMACS (Gera & Balasubramanian, 2020), ResNet101+BLSTM (Liu et al., 2020), ResNet101+BLSTM+CBAM (Liu et al., 2020), SIU (Dresvyanskiy et al., 2020), and TNT (Kuhnke et al., 2020).

Table 8 gives the performance comparisons between the proposed method and several state-of-the-art FER methods on in-the-lab databases (CK+, MMI, and Oulu-CASIA). Tables 9, 10, and 11 give the performance comparisons on two in-the-wild databases (RAF-DB and SFEW), Aff-Wild2, and AffectNet, respectively. The accuracy obtained by each competing method is taken directly from the corresponding paper.

4.4.1 Results on In-the-Lab Databases

As shown in Table 8, almost all the methods obtain high recognition accuracy in the CK+ database and relatively low classification rates in the MMI database among the three in-the-lab databases. This is because the images from CK+ are of high quality and the intensities of different expressions are strong, while those from MMI are affected by the glasses and the expression intensities are weak.

Among all the competing methods, the top four methods are our proposed ADDL, DDL, FN2EN, and PHRNN-MSCNN. The proposed ADDL method outperforms DDL in all the in-the-lab databases due to the effectiveness of ADFL and MINE, where the ADFL extracts adaptive disturbance-specific features and the MINE performs explicit disentanglement between expression-specific features and

Table 9 Performance comparisons between our method and several state-of-the-art FER methods on the RAF-DB and SFEW databases

Methods	Accuracy (%)	
	RAF-DB	SFEW
IACNN (Meng et al., 2017)	–	50.98
DLP-CNN (Li et al., 2017)	84.13	51.05
gACNN (Li et al., 2018)	85.07	–
IPA2LT (Zeng et al., 2018)	86.77	58.29
SPDNet (Acharya et al., 2018)	87.00	58.14
IPFR (Wang et al., 2019)	–	57.40
DAM-CNN (Xie et al., 2019a)	–	42.30
RAN (Wang et al., 2020c)	86.90	56.40
SCN** (Wang et al., 2020b)	88.14	–
DDL (Ruan et al., 2020)	87.71	59.86
PSR (Vo et al., 2020)	88.98	–
DACL (Farzaneh & Qi, 2021)	87.78	–
ADDL (proposed)	**89.34**	**62.16**

The best results are boldfaced. ** denotes that the RAF-DB and AffectNet are jointly used for training

disturbance-specific features. Note that the disturbing factors are not explicitly disentangled in DDL, leading to inferior expression-specific features. ADDL also achieves better accuracy than FN2EN on both CK+ and Oulu-CASIA. Note that our test set in CK+ is more challenging (since it contains the images corresponding to the contempt expression apart from the six basic expressions), while FE2EN only considers the six basic expressions. PHRNN-MSCNN is comprised of a recurrent neural network (RNN) and a CNN, where both the facial image and facial landmarks are used as the input. In contrast, our proposed ADDL achieves better performance by using a single image as the input. In particular, although MMI contains more challenging variations than the other two in-the-lab databases, ADDL outperforms PHRNN-MSCNN by a large margin (4.95% improvements) on MMI. This can be ascribed to the effectiveness of our proposed adaptive deep disturbance-disentangled learning.

Table 11 Performance comparisons between our method and several FER state-of-the-art methods on the AffectNet database

Methods	Accuracy (%)
IPA2LT (Zeng et al., 2018)	57.31
gACNN (Li et al., 2018)	58.78
IPFR (Wang et al., 2019)	57.40
FMPN (Chen et al., 2019)	61.52
CPG (Hung et al., 2019b)	63.57
PAENet (Hung et al., 2019a)	65.29
PSR (Vo et al., 2020)	63.77
DACL (Farzaneh & Qi, 2021)	65.20
EfficientNet-B0 (Savchenko, 2021)	65.74
ADDL (proposed)	**66.20**

The best results are boldfaced

4.4.2 Results on In-the-Wild Databases

As shown in Table 9, we compare the proposed method with twelve state-of-the-art FER methods on the RAF-DB and SFEW databases. Among all the methods, the proposed ADDL, SCN, DACL, DDL, and SPDNet obtain higher recognition accuracy than the other competing methods on RAF-DB, while the proposed ADDL, DDL, IPA2LT, and SPDNet are the top four methods on SFEW. SCN addresses the uncertainty problem in FER and achieves state-of-the-art performance. IPA2LT deals with the problem of inconsistent annotations in the FER databases. SPDNet introduces the covariance pooling into FER. PSR develops a scaling block to handle facial images at different resolutions. DACL lever-

Table 10 Performance comparisons between our method and several state-of-the-art FER methods on the Aff-Wild2 validation set

Methods	Input	F1 score (%)	Accuracy (%)	Overall (%)
CNN (Anas et al., 2020)	Image	29.16	50.77	36.29
NISL (Deng et al., 2020)	Image	–	–	42.43
LLAM (Wang et al., 2020a)	Image	38.00	49.00	42.00
ICT-VIPL (Zhang et al., 2020c)	Video&Audio	33.30	64.00	43.40
DMACS (Gera & Balasubramanian, 2020)	Image	37.00	**64.90**	46.50
ResNet101+BLSTM (Liu et al., 2020)	Video	28.10	64.70	40.20
ResNet101+BLSTM+CBAM (Liu et al., 2020)	Video	33.30	64.00	43.40
SIU (Dresvyanskiy et al., 2020)	Video&Audio	–	–	**56.56**
TNT (Kuhnke et al., 2020)	Video&Audio	–	–	54.60
ADDL (proposed)	Image	**42.23**	64.73	49.66

The best results are boldfaced

ages an attention mechanism based on a sparse center loss to enhance the discriminative capability of features. However, the above methods do not explicitly take the disturbing factors into consideration, which may lead to inferior performance in the case of limited training samples.

On the one hand, IACNN and IPFR are disturbance-disentangled-based methods, but they can cope with only one or two disturbing factors. Different from the above methods, ADDL is able to explicitly disentangle multiple disturbing factors by leveraging adversarial transfer learning, even though disturbing factors are not labeled in the FER database. On the other hand, gACNN and RAN address the occlusion problem by combining local learning and global learning. However, these two methods only utilize high-level features to perform FER. Unlike these methods, ADDL exploits both high-level features and low-level features in the expression subnetwork, thereby achieving excellent performance. Finally, compared with DDL, ADDL achieves higher accuracy on RAF-DB and SFEW. It is worth pointing out that DDL cannot adaptively choose the disturbing factors when trained on an FER database. However, ADDL effectively alleviates this problem by designing the ADFL.

From Table 10, our proposed ADDL method performs the best among all the image-based and video-based methods, with an overall score of 49.66%. SIU and TNT outperform our method, because they exploit the additional temporal and audio information for FER. Among all the competing methods, NISL proposes a multi-task model to learn from incomplete labels. LLAM and DMACS resort to attention blocks to extract global and local attention-aware features from facial images. ICT-VIPL, SIU, and TNT simultaneously extract visual features from videos and acoustic features from audio tracks to construct discriminative expression features. ResNet101+BLSTM uses ResNet-101 and BLSTM to extract semantic features and temporal features, respectively. However, the above methods do not fully consider the multiple disturbing factors in facial expression images. In summary, the above results show the effectiveness of our method in the large-scale FER database.

From Table 11, the proposed ADDL outperforms the other competing methods on AffectNet. FMPN designs an additional branch to learn local features from facial muscle moving regions. Then, the local features are combined with holistic features for classifying expressions. CPG and PAENet introduce compact and unforgetting models to progressively learn new tasks. EfficientNet-B0 is trained in a multi-task learning manner, where facial attribute prediction is performed to improve the representation ability of the features (i.e., edges and corners) at the lower CNN layers. However, the above methods along with IPA2LT and gACNN do not perform disturbance disentanglement, leading to inferior FER performance.

Table 12 The number of parameters and FLOPs obtained by different methods on the RAF-DB database

Methods	Training		
	Modules	Params	FLOPs
SCN	ResNet-18	11.2M	1.82G
Baseline_mat	S_g+S_e	16.2M	2.82G
	DFEM	11.4M	1.82G
ADDL	MINE	16.5K	16.4K
	ADDM (S_g+S_e+S_d)	20.6M	3.82G

4.5 Computational Complexity

In this subsection, we briefly analyze the computational complexity of the ADDL method. We also evaluate SCN and the Baseline_mat method for a comparison. Note that the results obtained by other competing methods are not given since their source codes are not publicly available. We use the number of parameters (Params) and Floating Point operations (FLOPs) to evaluate the memory consumption and computational complexity of the model, respectively. Moreover, we adopt the inference time and speed to measure latency. We take the RAF-DB database for performance evaluation.

Table 12 reports the number of parameters and FLOPs obtained by SCN, Baseline_mat, and ADDL. Both ADDL and Baseline_mat have more parameters and higher FLOPs than SCN. This is because the ADDM, which is based on multiple attention blocks, is trained during the two-stage learning procedure.

The inference time and speed obtained by SCN, Baseline_mat, and ADDL are given in Table 13. We can observe that the proposed ADDL obtains an inference time of 5.21 ms, which is similar to Baseline_mat due to the same inference phases. The inference speed of ADDL is lower than that of SCN. Because multiple attention blocks are employed in ADDL to extract discriminative features. This improves the FER accuracy but slows down the inference speed of the model. Although the computational complexity of the training phase of our proposed ADDL method is high, it can still obtain real-time inference speed and be applicable to real-world scenarios.

4.6 Valence and Arousal Estimation

In this subsection, we evaluate the performance of our method for the task of valence and arousal (VA) estimation on the AffectNet database. Similar to previous methods (Mollahos-

Table 13 The inference time and speed obtained by different methods on the RAF-DB database

Methods	Testing	
	Inference time (ms)	Speed (FPS)
SCN	3.72	268.88
Baseline_mat	5.17	193.52
ADDL	5.21	192.12

The inference time and speed are measured in milliseconds (ms) and frames per second (FPS), respectively

seini et al. 2017; Jang et al., 2019; Kollias et al., 2018), we view the VA estimation as a regression task.

To perform the VA estimation, an FC layer is added after the expression classification layer (i.e., the last layer) in S_e to regress the valence and arousal values. Then, we fine-tune the classification and regression layers based on a well-trained ADDM that obtains the best validation accuracy for a 7-way FER. The learning rate is set to 0.001 for the last two layers and 0.0001 for the other layers in the ADDM. In this paper, we adopt two commonly used metrics, i.e., root mean square error (RMSE) and concordance correlation coefficient (CCC) (Mollahosseini et al. 2017), to evaluate the performance. Thus, we add the RMSE and CCC losses into Eq. (14) for joint training. The comparison results are reported in Table 14.

As shown in Table 14, the factorized high-order CNN method achieves the best performance on the four evaluation metrics except for the RMSE of valence. The proposed ADDL obtains the best result on the RMSE of valence and the second place on the other three evaluation metrics. Factorized high-order CNN (Kossaifi et al., 2020b) employs a higher-order factorized convolution network, where a single tensor regression layer (Kossaifi et al., 2020a) is dedicated to performing regression of the VA values. In contrast, the proposed ADDL is based on a classification model with an additional regression layer, which may limit the regression performance. The VGG-Face+2M imgs method synthesizes facial images to improve the performance for the VA estimation. Face-SSD jointly performs face detection and face analysis. However, these methods obtain worse performance than ours. These results show the feasibility of our method for the VA estimation.

5 Conclusion and Future Work

In this paper, we propose a novel ADDL method for FER. ADDL is able to disentangle multiple disturbing factors simultaneously and adaptively (even when the labels of disturbing factors are not available in the FER database) and effectively extract expression-related information. The training of ADDL contains two stages. First, a DFEM is trained to identify multiple disturbing factors in a multi-task learning manner. Then, based on the trained DFEM, an ADDM is learned to classify facial expressions by considering the characteristics of the FER database. In the ADDM, an ADFL is developed to estimate the importance weights corresponding to different disturbing factors and perform adversarial transfer learning. Furthermore, an MINE is employed to enable the explicit disentanglement between expression-specific features and disturbance-specific features. Extensive experiments on both in-the-lab and in-the-wild FER databases have demonstrated the superior performance of ADDL over several state-of-the-art FER methods.

It is widely assumed that facial expressions can infer the emotional state of humans. However, Barrett et al. (2019) show that the way humans express their emotions may significantly vary across different cultures and situations. Moreover, they also reveal that similar configurations of facial movements may belong to different emotion categories. Naturally, human perception of emotions does not rely on one type of information. Instead, it is triggered by a variety of cues from different sources. By investigating such cues, many recent efforts (Lv et al., 2021) have been proposed toward multi-modality (such as facial expressions, body gestures, and voice to physiological signals) emotion recognition by leveraging the strengths of each modality. In the future, we plan to extend our method to multi-modality emotion recognition.

Table 14 The results of valence and arousal estimation on the AffectNet databas

Methods	Valence		Arousal	
	CCC	RMSE	CCC	RMSE
AlexNet (Mollahosseini et al. 2017)	0.60	0.37	0.34	0.41
Face-SSD (Jang et al., 2019)	0.57	0.44	0.47	0.39
VGG-Face+2M imgs (Kollias et al., 2018)	0.62	0.37	0.54	0.39
Factorized higher-order CNN (Kossaifi et al., 2020b)	**0.71**	0.35	**0.63**	**0.32**
ADDL (proposed)	0.66	**0.34**	0.59	0.33

The best results are boldfaced

Acknowledgements This work was partly supported by the National Natural Science Foundation of China under Grants 62071404, U21A20514, and 61872307, by the Open Research Projects of Zhejiang Lab under Grant 2021KG0AB02, by the Natural Science Foundation of Fujian Province under Grant 2020J01001, and by the Youth Innovation Foundation of Xiamen City under Grant 3502Z20206046.

References

Acharya, D., Huang, Z., Pani Paudel, D., & Van Gool, L. (2018). Covariance pooling for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 367–374).

Anas, H., Rehman, B., & Ong, W. H. (2020) Deep convolutional neural network based facial expression recognition in the wild. arXiv preprint arXiv:2010.01301

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in The Public Interest, 20*(1), 1–68.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., & Hjelm, D. (2018). Mutual information neural estimation. In *Proceedings of the International conference on machine learning* (pp. 531–540).

Chang, F. J., Tran, A. T., Hassner, T., Masi, I., Nevatia, R., & Medioni, G. (2019). Deep, landmark-free fame: Face alignment, modeling, and expression estimation. *International Journal of Computer Vision, 127*(6–7), 930–956.

Chen, J., Konrad, J., & Ishwar, P. (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* (pp. 1570–1579).

Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X., & Rui, Y. (2020). Label distribution learning on auxiliary label space graphs for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 13984–13993).

Chen, Y., Wang, J., Chen, S., Shi, Z., & Cai, J. (2019) Facial motion prior networks for facial expression recognition. In *Proceedings of the IEEE conference on visual communications and image processing* (pp. 1–4).

Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A. L., & Wang, X. (2017). Multi-context attention for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1831–1840).

Dapogny, A., Bailly, K., & Dubuisson, S. (2018). Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. *International Journal of Computer Vision, 126*(2–4), 255–271.

Deng, D., Chen, Z., & Shi, B. E. (2020) Multitask emotion recognition with incomplete labels. In *Proceedings of IEEE international conference on automatic face & gesture recognition* (pp. 828–835).

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2011). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 2106–2112).

Ding, H., Zhou, S. K., & Chellappa, R. (2017). FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition. In *Proceedings of the international conference on automatic face and gesture recognition* (pp. 118–126).

Donsker, M. D., & Varadhan, S. S. (1983). Asymptotic evaluation of certain Markov process expectations for large time. IV. *Communications on Pure and Applied Mathematics, 36*(2), 183–212.

Dresvyanskiy, D., Ryumina, E., Kaya, H., Markitantov, M., Karpov, A., & Minker, W. (2020) An audio-video deep and transfer learning framework for multimodal emotion recognition in the wild. arXiv preprint arXiv:2010.03692

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental Psychology and Nonverbal Behavior, 1*(1), 56–75.

Farzaneh, A. H., & Qi, X. (2021) Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of IEEE winter conference on applications of computer vision* (pp. 2402–2411).

Fu, J., Zheng, H., & Mei, T. (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* (pp. 4438–4446).

Gera, D., & Balasubramanian, S. (2020) Affect expression behaviour analysis in the wild using spatio-channel attention and complementary context information. arXiv preprint arXiv:2009.14440

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multipie. *Image and Vision Computing, 28*(5), 807–813.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Hu, P., Cai, D., Wang, S., Yao, A., & Chen, Y. (2017). Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of ACM international conference on multimodal interaction* (pp. 553–560).

Hung, S. C., Lee, J. H., Wan, T. S., Chen, C. H., Chan, Y. M., & Chen, C. S. (2019a) Increasingly packing multiple facial-informatics modules in a unified deep-learning model via lifelong learning. In *Proceedings of the international conference on multimedia retrieval* (pp. 339–343).

Hung, S. C., Tu, C. H., Wu, C. E., Chen, C. H., Chan, Y. M., & Chen, C. S. (2019b) Compacting, picking and growing for unforgetting continual learning. arXiv preprint arXiv:1910.06562

Jang, Y., Gunes, H., & Patras, I. (2019). Registration-free face-SSD: Single shot analysis of smiles, facial attributes, and affect in the wild. *Computer Vision and Image Understanding, 182,* 17–29.

Jung, H., Lee, S., Yim, J., Park, S., & Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 2983–2991).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kollias, D., & Zafeiriou, S. (2018) Aff-Wild2: Extending the aff-Wild database for affect recognition. arXiv preprint arXiv:1811.07770

Kollias, D., Cheng, S., Ververas, E., Kotsia, I., & Zafeiriou, S. (2018) Generating faces for affect analysis. arXiv preprint arXiv:1811.05027

Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B., et al. (2019). Deep affect prediction in-the-wild: Aff-Wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision, 127*(6), 907–929.

Kollias, D., Cheng, S., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020a). Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision, 128*(5), 1455–1484.

Kollias, D., Schulc, A., Hajiyev, E., & Zafeiriou, S. (2020b) Analysing affective behavior in the first ABAW 2020 competition. arXiv preprint arXiv:2001.11409

Kossaifi, J., Lipton, Z. C., Kolbeinsson, A., Khanna, A., Furlanello, T., & Anandkumar, A. (2020a). Tensor regression networks. *Journal of Machine Learning Research, 21,* 1–21.

Kossaifi, J., Toisoul, A., Bulat, A., Panagakis, Y., Hospedales, T. M., & Pantic, M. (2020b) Factorized higher-order CNNs with an application to spatio-temporal emotion estimation. In *Proceedings of* *the IEEE conference on computer vision and pattern recognition* (pp. 6060–6069).

Kuhnke, F., Rumberg, L., & Ostermann, J. (2020). Two-stream aural-visual affect analysis in the wild. arXiv preprint arXiv:2002.03399

Li, S., & Deng, W. (2018). Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing, 28*(1), 356–370.

Li, S., & Deng, W. (2019). Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision, 127*(6–7), 884–906.

Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 1–25.

Li, S., Deng, W., & Du, J. (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* (pp. 2852–2861).

Li, Y., Zeng, J., Shan, S., & Chen, X. (2018). Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing, 28*(5), 2439–2450.

Li, Y., Zeng, J., Shan, S., & Chen, X. (2019) Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 10924–10933).

Liu, H., Zeng, J., Shan, S., & Chen, X. (2020) Emotion recognition for in-the-wild videos. arXiv preprint arXiv:2002.05447

Liu, S., Johns, E., & Davison, A. J. (2019). End-to-end multi-task learning with attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1871–1880).

Liu, Y., Zeng, J., Shan, S., & Zheng, Z. (2018). Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition. In *Proceedings of the international conference on automatic face and gesture recognition* (pp. 458–465).

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015) Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision* (pp. 3730–3738).

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (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* (pp. 94–101).

Lv, F., Chen, X., Huang, Y., Duan, L., & Lin, G. (2021) Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2554–2562).

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(11), 2579–2605.

Meng, Z., Liu, P., Cai, J., Han, S., & Tong, Y. (2017). Identity-aware convolutional neural network for facial expression recognition. In *Proceedings of the international conference on automatic face and gesture recognition* (pp. 558–565).

Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. In *Proceedings of the IEEE winter conference on applications of computer vision* (pp. 1–10).

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10*(1), 18–31.

Motley, M. T., & Camden, C. T. (1988). Facial expression of emotion: A comparison of posed expressions versus spontaneous expressions in an interpersonal communication setting. *Western Journal of Communication (includes Communication Reports), 52*(1), 1–22.

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

Rifai, S., Bengio, Y., Courville, A., Vincent, P., & Mirza, M. (2012). Disentangling factors of variation for facial expression recognition. In *Proceedings of the European conference on computer vision* (pp. 808–822).

Ruan, D., Yan, Y., Chen, S., Xue, J-H., & Wang, H. (2020). Deep disturbance-disentangled learning for facial expression recognition. In *Proceedings of the ACM international conference on multimedia* (pp. 2833–2841).

Sankaran, N., Mohan, D. D., Lakshminarayana, N. N., Setlur, S., & Govindaraju, V. (2020). Domain adaptive representation learning for facial action unit recognition. *Pattern Recognition, 102,* 107–127.

Savchenko, A. V. (2021) Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. arXiv preprint arXiv:2103.17107

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Valstar, M., & Pantic, M. (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* (pp. 65–70).

Vo, T. H., Lee, G. S., Yang, H. J., & Kim, S. H. (2020). Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access, 8,* 131988–132001.

Wang, C., Wang, S., & Liang, G. (2019). Identity- and pose-robust facial expression recognition through adversarial feature learning. In *Proceedings of the ACM international conference on multimedia* (pp. 238–246).

Wang, C., Hu, R., Hu, M., Liu, J., Ren, T., He, S., Jiang, M., & Miao, J. (2020a) Lossless attention in convolutional networks for facial expression recognition in the wild. arXiv preprint arXiv:2001.11869

Wang, K., Peng, X., Yang, J., Lu, S., & Qiao, Y. (2020b). Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6897–6906).

Wang, K., Peng, X., Yang, J., Meng, D., & Qiao, Y. (2020c). Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing, 29*(1), 4057–4069.

Wang, W., Fu, Y., Sun, Q., Chen, T., Cao, C., Zheng, Z., Xu, G., Qiu, H., Jiang, Y., & Xue, X. (2020d). Learning to augment expressions for few-shot fine-grained facial expression recognition. arXiv preprint arXiv:2001.06144

Wu, L., Wang, Y., Gao, J., & Li, X. (2018). Where-and-when to look: Deep siamese attention networks for video-based person re-identification. *IEEE Transactions on Multimedia, 21*(6), 1412–1424.

Xie, S., Hu, H., & Wu, Y. (2019a). Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. *Pattern Recognition, 92,* 177–191.

Xie, S., Hu, H., & Chen, Y. (2020). Facial expression recognition with two-branch disentangled generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology, 31*(6), 2359–2371.

Xie, W., Jia, X., Shen, L., & Yang, M. (2019b). Sparse deep feature learning for facial expression recognition. *Pattern Recognition, 96,* 106966.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International conference on machine learning* (pp. 2048–2057).

Yan, Y., Huang, Y., Chen, S., Shen, C., & Wang, H. (2020). Joint deep learning of facial expression synthesis and recognition. *IEEE Transactions on Multimedia, 22*(11), 2792–2807.

Yang, H., Ciftci, U., & Yin, L. (2018a). Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2168–2177).

Yang, H., Zhang, Z., & Yin, L. (2018b). Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In *Proceedings of the International conference on automatic face and gesture recognition* (pp. 294–301).

Yu, Z., & Zhang, C. (2015). Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the ACM international conference on multimodal interaction* (pp. 435–442).

Zeng, J., Shan, S., & Chen, X. (2018). Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision* (pp. 222–237).

Zhang, F., Zhang, T., Mao, Q., Duan, L., & Xu, C. (2018a). Facial expression recognition in the wild: A cycle-consistent adversarial attention transfer approach. In *Proceedings of the ACM international conference on multimedia* (pp. 126–135).

Zhang, F., Zhang, T., Mao, Q., & Xu, C. (2018b). Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3359–3368).

Zhang, F., Zhang, T., Mao, Q., & Xu, C. (2020a). Geometry guided pose-invariant facial expression recognition. *IEEE Transactions on Image Processing, 29,* 4445–4460.

Zhang, F., Zhang, T., Mao, Q., & Xu, C. (2020b). A unified deep model for joint facial expression recognition, face synthesis, and face alignment. *IEEE Transactions on Image Processing, 29,* 6574–6589.

Zhang, K., Huang, Y., Du, Y., & Wang, L. (2017). Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing, 26*(9), 4193–4203.

Zhang, Y. H., Huang, R., Zeng, J., Shan, S., & Chen, X. (2020c) M^{3}T: Multi-modal continuous valence-arousal estimation in the wild. arXiv preprint arXiv:2002.02957

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2018c). From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision, 126*(5), 550–569.

Zhang, Z., Zhai, S., & Yin, L. (2018d) Identity-based adversarial training of deep CNNS for facial action unit recognition. In *Proceedings of the British machine vision conference* (pp. 1–13).

Zhao, G., & Pietikäinen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(6), 915–928.

Zhao, G., Huang, X., Taini, M., Li, S. Z., & PietikäInen, M. (2011). Facial expression recognition from near-infrared videos. *Image and Vision Computing, 29*(9), 607–619.

Zhao, T., & Wu, X. (2019). Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3085–3094).

Zhao, X., Liang, X., Liu, L., Li, T., Han, Y., Vasconcelos, N., & Yan, S. (2016). Peak-piloted deep network for facial expression recognition. In *Proceedings of the European conference on computer vision* (pp. 425–442).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.