International Journal of Computer Vision (2019) 127:884–906
https://doi.org/10.1007/s11263-018-1131-1

Blended Emotion in-the-Wild: Multi-label Facial Expression Recognition Using Crowdsourced Annotations and Deep Locality Feature Learning

Shan Li[1] · **Weihong Deng**[1]

Received: 15 January 2018 / Accepted: 13 November 2018 / Published online: 29 November 2018

Abstract
Comprehending different categories of facial expressions plays a great role in the design of computational model analyzing human perceived and affective state. Authoritative studies have revealed that facial expressions in human daily life are in multiple or co-occurring mental states. However, due to the lack of valid datasets, most previous studies are still restricted to basic emotions with single label. In this paper, we present a novel multi-label facial expression database, RAF-ML, along with a new deep learning algorithm, to address this problem. Specifically, a crowdsourcing annotation of 1.2 million labels from 315 participants was implemented to identify the multi-label expressions collected from social network, then EM algorithm was designed to filter out unreliable labels. For all we know, RAF-ML is the first database in the wild that provides with crowdsourced cognition for multi-label expressions. Focusing on the ambiguity and continuity of blended expressions, we propose a new deep manifold learning network, called Deep Bi-Manifold CNN, to learn the discriminative feature for multi-label expressions by jointly preserving the local affinity of deep features and the manifold structures of emotion labels. Furthermore, a deep domain adaption method is leveraged to extend the deep manifold features learned from RAF-ML to other expression databases under various imaging conditions and cultures. Extensive experiments on the RAF-ML and other diverse databases (JAFFE, CK+, SFEW and MMI) show that the deep manifold feature is not only superior in multi-label expression recognition in the wild, but also captures the elemental and generic components that are effective for a wide range of expression recognition tasks.

Keywords Facial expression recognition · Deep feature learning · Multi-label classification · Crowdsourced database in-the-wild

1 Introduction

Automatic facial expression analysis has attracted broad attention due to its numerous potential applications in social media analysis and human–computer interaction (HCI). Research on recognizing emotion through facial expressions can be traced back to the 1970s, when Ekman and Rosenberg (1997) have defined six basic expressions on account of extensive studies, namely happiness, sadness, anger, surprise, disgust and fear. Because of their evolutionary significance, most of the previous studies on the discrete categorical emotion description stream have regarded the emotion recognition problem as a singular classification problem and turned to classifying expressions into one of those six categories or seven categories (plus neutral). It is reported that nearly up to 100% performances has been achieved on this single-label classification task (Li and Deng 2018), e.g., 98.9% obtained in Zhang et al. (2018).

However, while analyzing natural human interactions, one cannot expect that every human will express clear emotional content. Authoritative studies Ekman and Friesen (2003),

Communicated by Rama Chellappa, Xiaoming Liu, Tae-Kyun Kim, Fernando De la Torre, Chen Change Loy.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11263-018-1131-1) contains supplementary material, which is available to authorized users.

✉ Weihong Deng
whdeng@bupt.edu.cn

Shan Li
ls1995@bupt.edu.cn

[1] Beijing University of Posts and Telecommunications (BUPT), Beijing, China

Hassin et al. (2013), Izard (1972) and Plutchik (1991) have discovered that, humans' facial representations are often not pure examples of a single expression category, but always admixtures of different emotions; that is, they appear as combinations, blends, or compounds of different basic emotions. For example, one may feel both surprised and fearful concurrently when getting an unexpected fright. It is grounded that treating facial expression analysis as a single-label classification problem is over simplified and lack of practical applicability.

One line of approaches (Ding et al. 2013; Eleftheriadis et al. 2015b; Pantic and Rothkrantz 2000; Zeng et al. 2015) to this complex emotion problem are based on facial action coding system (FACS) analysis (Ekman and Rosenberg 1997). Unfortunately, annotating large number of action units (AUs) of facial images, especially in unconstrained real-word conditions, is extremely time-consuming and requires professionally trained annotators. And it is also quite difficult to correspond the combinations of AUs to specific emotion category.

In contrast, the other line of approaches based on *multi-label expression recognition* is more convenient and accessible, which can intuitively figure out blended expressions and quantify the recognition result into multiple emotion labels. However, related studies Chang et al. (2004), Wang et al. (2014), Zhou et al. (2015) and Zhao et al. (2015) were conducted on few small-scale lab-controlled databases, which may not be applicable to our daily life. Limitation in this case is mainly caused by the lack of specific databases. Detailed investigations on facial expression databases have showed that most of the existing databases only provide samples attached to one label, i.e., single emotion label is associated with each instance. There have been no facial expression databases, as far as we know, that explicitly consider humans' various perception of emotion both in the image collection and annotation process.

To address this issue, we treat the recognition task as a multi-label problem and construct a novel database, *Real-world Affective Face-Multi Label* (*RAF-ML*),[1] for multi-label expression analysis. First of all, a great amount of facial images in different occlusions, illuminations and resolutions from thousands of different individuals were collected from the social network. To pick out images with blended expressions, motivated by Ekman's theory,[2] we employed 315 well-trained annotators to ensure each image can be annotated enough independent times. Furthermore, an EM based reliability evaluation algorithm is proposed to get a reliable emotion probability vector for each image. By analyzing 1.2 million labels of around thirty thousand unconstrained images, 4908 images with multi-peak label distribution have been selected out to constitute RAF-ML.[3] To our best knowledge, RAF-ML is the first in-the-wild facial expression database that is specially designed for affective images attached with certain multiple tags via crowdsourced annotation. Figure 1 exhibits examples of samples in RAF-ML that presented with multi-label expressions under various real-world conditions.

Deep learning has been the state-of-the-art technique on many unconstrained recognition tasks, however, there has been no suitable model focusing on the subtlety, complexity and continuity of multi-label expressions in the wild. To achieve this goal, we propose a new deep manifold feature learning based framework, *Deep Bi-Manifold CNN* (*DBM-CNN*), which simultaneously and efficiently considers crowd-sourced label information and feature compactness in the low-dimensional manifolds by adding a new loss layer, bi-manifold loss. Jointly trained with the cross-entropy loss which forces images with different labels to stay apart, the bi-manifold loss drives the locally neighboring faces sharing the similar intensity distribution to become coherent and thus the discriminative power of the deeply learned features can be highly enhanced.

To enhance the generalization ability of the learned features, we extend the DBM-CNN to learn more transferable representations for other related expression databases by embedding domain adaptation in the pipeline of deep learning. As DBM-CNN learns deep features that eventually transition from general to specific along the network, a multi-kernel maximum mean discrepancies (MK-MMD) loss layer is appended to DBM-CNN to align the statistical distribution shift between the training set RAF-ML and other test datasets. Moreover, for the special scenario that transfers from the multi-label source domain to the single-label target domain, an entropy loss is further employed to force the learned label space on the target data to present an unimodal distribution, so that the output probability vectors are closer to the single-label ground truth. By matching the mean embeddings of these two different domains, the capacity of preserving the local clusters on bi-manifolds can thus be efficiently applied to a great diversity of facial expression recognition (FER) scenarios.

Extensive experiments on RAF-ML were conducted by comparing several widely-used handcrafted features and deep learning features across different multi-label classification algorithms and various evaluation metrics. Results

[1] http://www.whdeng.cn/RAF/model2.html.

[2] Ekman have stated in Ekman et al. (2013), "if the stimulus does contain an emotion blend, and the investigator allows only a single choice which does not contain blend terms, low levels of agreement may result, since some of the observers may choose a term for one of the blend components, some for another."

[3] Compound emotions out of these 4908 images and the other basic emotions have been presented in RAF-DB (Li et al. 2017; Li and Deng 2019).

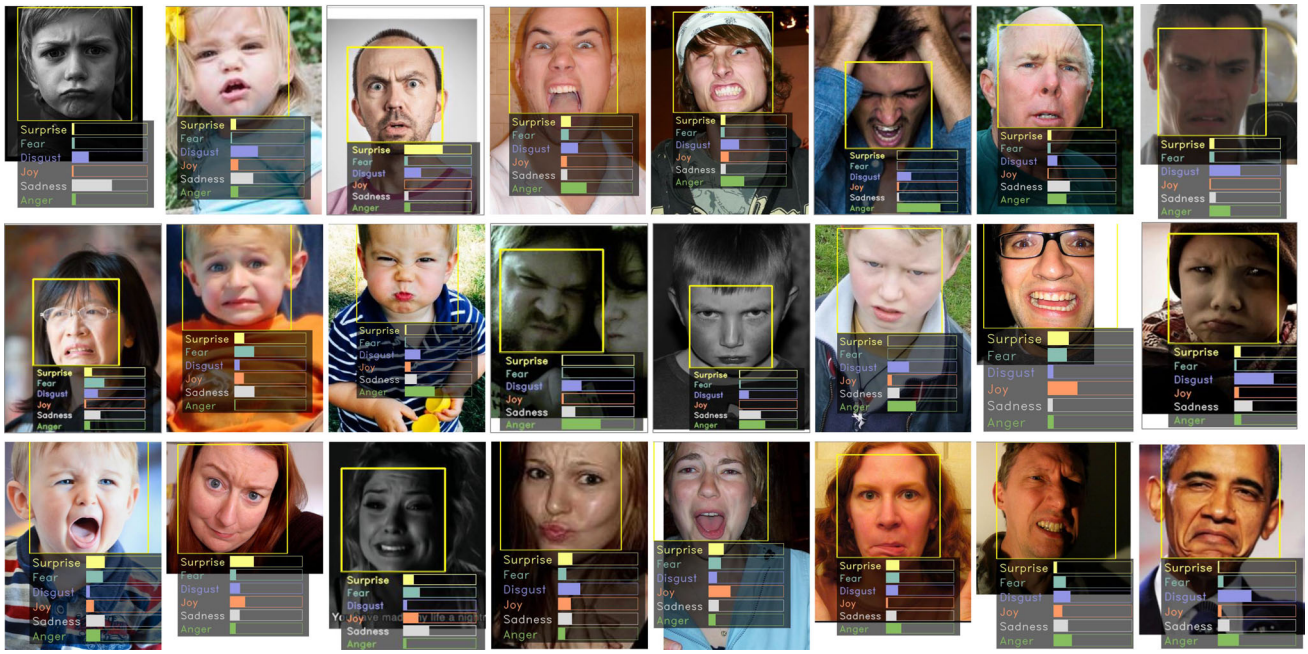


Fig. 1 Examples with multi-label emotions from RAF-ML. We can see that people in our daily life tend to express a combination of several different repertoire of feelings. Therefore, it is misguided to simply describe human emotion and affect using only one single component expression

show that our deep manifold feature outperforms both handcrafted and other state-of-the-art CNN features. Moreover, by implementing domain adaption for DBM-CNN, the activation features trained on RAF-ML can generalize well to other unseen databases, such as multi-label database JAFFE (Lyons et al. 1998) and single-label databases CK+ (Lucey et al. 2010), SFEW 2.0 (Dhall et al. 2015b) and MMI (Valstar and Pantic 2010), which suggests that our network can generate universal features to handle a wide range of emotion analysis tasks for different expression datasets.

The reminder of this paper is organized as follows. Section 2 discusses the definition, existing datasets, and the state-of-the art learning methods for multi-label FER, and also reviews the related approaches for facial expression domain adaption. Section 3 presents details of the construction process of RAF-ML. Section 4 proposes the new deep manifold feature learning model DBM-CNN. Section 5 introduces the deep domain adaption method to transfer the knowledge gained from RAF-ML to other related expression databases. In Sect. 6, we include the experimental results of different features on RAF-ML. Then we evaluate the generalizability of our method on different FER tasks. Finally, we discuss and conclude the work in Sect. 7.

2 Related Work

In this section, we first investigate the definition of multi-label facial expressions and relevant existing databases. Then we review the related work in multi-label FER and domain adaption methods for FER.

2.1 Multi-label Expression Definition

Since the twentieth century, numerous psychological studies and cognitive sciences have endorsed the theory that the capacity of the face frequently contains components of more than one emotion at the given instant, observable even in still facial photographs.

Tomkins (1963) discussed how various emotions come to be combined. The author gave an example that a child may experience an emotional state that is a mixture of fear and shame when under certain patterns of parenting. In Ekman and Scherer (1984), the authors' studies of self-report suggested that people typically experience blends of emotions. If subjects are asked to imagine fear they are likely to generate fear blended with surprise or distress. Experiments conducted by Nummenmaa (1988) certified that it is possible for human to express pleasure, surprise, hate, fear, and sorrow, and pairwise combinations of these. Still photographs are also proved to be useful for this specific purpose. Izard (2013) also listed some common patterns or combinations of affects. For example, anxiety can be defined as a mixture of sadness, fear and anger. Individual variations in the patterns of basic emotions can yield different kinds of anxiety. Most recently, Du et al. (2014) have proposed compound facial expressions that are constructed as a combination of two basic emotion categories and identified 15 compound expressions consistently

produced across cultures. Their theories all indicate that the single discrete emotions can be blended or fused to form new emotions and these prototype emotions should prove useful in delineating the precise blended composition of mental representations.

Therefore, in this paper, in order to learn a practical facial expression analyzer which can intuitively figure out blended expressions in our daily life and quantify the recognize result into multiple emotion labels, we hinge on the theory of the blending of six basic emotions and model facial expression analysis as a multi-label classification problem.

2.2 Multi-label Expression Databases

According exhaustive investigations Anitha et al. (2010) and Sariyanidi et al. (2015) on facial expression databases, most existent databases only provide images attached with one expression category. And only two datasets provide expression data with multilabel information: the lab-controlled database JAFFE (Lyons et al. 1998) and the in-the-wild database EmotioNet (Fabian Benitez-Quiroz et al. 2016).

JAFFE database includes 213 facial images of only ten Japanese female subjects posing the six basic expressions plus neutral expression. Each of the subjects poses three to four examples per expression to make a total of 213 gray-scaled images in the size of 256×256 pixels. The images were captured under controlled environment in terms of pose and illumination, but it is worth mentioning that besides a single label representing the predominant expression of each image, semantic ratings of the expressions are provided as well, which represent the intensity on each emotion. A five-level scale was used for each of the 6 adjectives (5 represents highest emotion intensity, while 1 represents lowest emotion intensity).

EmotioNet is a large-scale database with one million facial expression images collected from the Internet by selecting all the words derived from the word "feeling" in WordNet. Most samples were annotated by an automatic AU detection algorithm, and the remaining 10% were manually annotated with AUs. EmotioNet contains 6 basic expressions plus neutral expression and also 17 compound expressions; however, the emotion categories are inferred from the AU labels without directly manual annotation and the multi-label expression categories it includes are composed of only two different basic component emotions.

For more thorough comparison, we also discuss the lab-controlled database BU-3DFE (Yin et al. 2006) and the in-the-wild database HAPPEI (Dhall et al. 2015a) which provide samples with single emotion label that attached with intensity information. The lab-controlled BU-3DFE database contains 606 facial expression sequences captured from 101 subjects who are requested to perform seven prototypic emotional states. Each sample is represented by one basic expression label with intensity information which are valued by the subjects and two psychologists. HAPPEI database consists 4886 images with multiple faces downloaded from Flickr. Without self-rating, in this database, labels are obtained from the perception of the annotators. And for these images, discrete levels of happiness intensity were manually annotated for 8500 faces. The face-level happiness intensity's range is [0–5], which corresponds to six stages of happiness.

For the two databases that created in lab-controlled environment, during the generation stage, participators in these two databases were requested to perform just one of the prototypic facial expressions deliberately. And images were produced in tightly controlled environments that are short of diversity on subjects and conditions. For the other two databases that collected from real world, there still exist some common deficiencies. The emotion intensity information for each example cannot cover all six basic expressions and the number of relevant labels to compose multi-label expression is limited. Moreover, the number of annotators is insufficient to guarantee the reliability and validity of the emotion labels. So in this paper, we propose a new database, RAF-ML, that provides images from various environments with multi-label expressions based on group perception and label data with minimal noise.

2.3 Multi-label Learning in Facial Expression Recognition

During the past decades, significant amount of progresses have been made on the multi-label classification learning paradigm. Detailed definition, evaluation metrics and representative multi-label learning algorithms can be seen in Tsoumakas and Katakis (2006) and Zhang and Zhou (2014). However, very few models of multi-label facial expressions have been developed so far for facial expression analysis.

In Chang et al. (2004), the transition between different expressions is presented as the evolution of the posterior probability of the six basic paths via a probabilistic model which can recognize blended expressions. In Wang et al. (2014), a novel approach of implicit multiple emotional video tagging is proposed which considers the relationship among the facial multi-expressions, and the relationships among the expression and emotions. In Zhao et al. (2015), multi-label Group Lasso regularized maximum margin classifier (GLMM) and Group Lasso regularized regression (GLR) algorithms are proposed which can model FER jointly with multiple outputs. In Zhou et al. (2015), an emotion distribution learning (EDL) algorithm has been proposed which learns the specific description degrees of all six basic emotions and maps the given expression image to the emotion distributions. In Xing et al. (2016), an additive weighted func-

tion regression from statistical viewpoint, Logistic Boosting Regression (LogitBoost), is used to constitute two Label Distribution Learning (LDL) algorithms named LDLogitBoost and AOSO-LDLogitBoost, which can lead to better performances on expression recognition.

Different from these methods that conducted on small-scaled laboratory-controlled facial expression databases and are shallow-learned, we proposed a novel deep network combined with manifold learning and evaluated our method on the in-the-wild database, RAF-ML.

Apart from the above mentioned methods conducted on multi-label facial expression, deep label distribution learning (DLDL) is proposed in Gao et al. (2017). Specifically, the discrete label distribution information is first constructed for various popular visual tasks under proper assumptions. Then, the Kullback–Leibler divergence between the predicted and obtained label distributions is minimized using deep CNNs to learn the label distribution. In our paper, we directly treat FER as a multi-label classification task and provide the ground-truth label distribution of training samples. We also employ the KL loss (i.e., multi-label softmax cross-entropy loss in this paper) to supervise the learning process and explore the label distribution information. Moreover, we propose a bi-manifold loss to further help enhance the discrimination ability of the learned deep features.

2.4 Domain Adaption in Facial Expression Recognition

Numerous approaches in the field of domain adaption have been proposed in the last years to address adaption problems that arise in different computer visual scenarios. However, very few significant interest until now has been gained in the cross-domain learning of facial expression recognition.

In Yan et al. (2011), the authors investigated the cross-dataset FER problem on facial expression and proposed a transfer subspace learning method. Afterwards, Miao et al. (2012) proposed a supervised extension of Kernel Mean Matching to match the distributions between different facial expression databases. In Chen et al. (2013), a person-specific model was proposed to transfer the informative knowledge from other people to a new subject with a small amount of training data. In Chu et al. (2013), a transductive learning method, Selective Transfer Machine (STM), was proposed to personalize a generic classifier by attenuating person-specific mismatches. And Zen et al. (2016) proposed a regression framework, Transductive Parameter Transfer (TPT), to build personalizing classification models which does not require labeled target data. In Zhu et al. (2016), a discriminative feature adaptation method was proposed to minimize the mismatch between different expression databases. Most recently, Zong et al. (2017) proposed Target Sample Re-Generator approach to re-generate samples sharing the same or similar distribution with the source ones for cross-domain micro-expression recognition.

In contrast to methods mentioned that transfer among different basic expression databases or different people within one single database, our approach makes an effort to adapt features learned from multi-label facial expressions database to basic expression recognition based on the deep learning technique.

3 RAF-ML

In this section, we go into details about the image collection, crowd-sourcing annotation and evaluation, multi-label construction and statistic metadata on RAF-ML dataset.

3.1 Collecting Process

To collect images with blended expressions, a set of emotion-related keywords combined with others related to age, race and gender were used as query words. And then image search engine of Flickr which has well-structured XML format search API with abundant user-added tags and much fewer duplicates than other search engines was queried with these keywords to download images containing naturalistic emotions in batches as many as possible.

After collecting sufficient affective images, a crowdsourcing annotation was used to select images with multi-label expressions. Naming and labeling emotions is a difficult and time-consuming task, especially when dealing with authentic data collected from challenging real-world conditions. In order to control the annotation difficulty, inspired by Ekman's theory (2013), we employed 315 annotators (students and staffs from universities) who have been instructed with a one-hour tutorial of psychological knowledge on emotion for an online annotation assignment, during which they were asked to classify images into the most apparent one from seven classes. During annotation, each image was labeled by about 40 independent labelers to ensure that multiple perceptions of images' affective state can be collected from sufficient observers. As a result, a multi-label annotation result is obtained for each image. Figure 2 shows the pipeline of data collection and annotation.

3.2 Reliability Estimation

Due to subjectivity and varied expertise of labelers and wide ranging levels of images' difficulty, there were some disagreements among annotators. Instead of employing the disambiguation methods with convex optimization (Cour et al. 2011; Zhang and Yu 2015) to get rid of noisy labels, we chose a parametric model, the Expectation Maximization

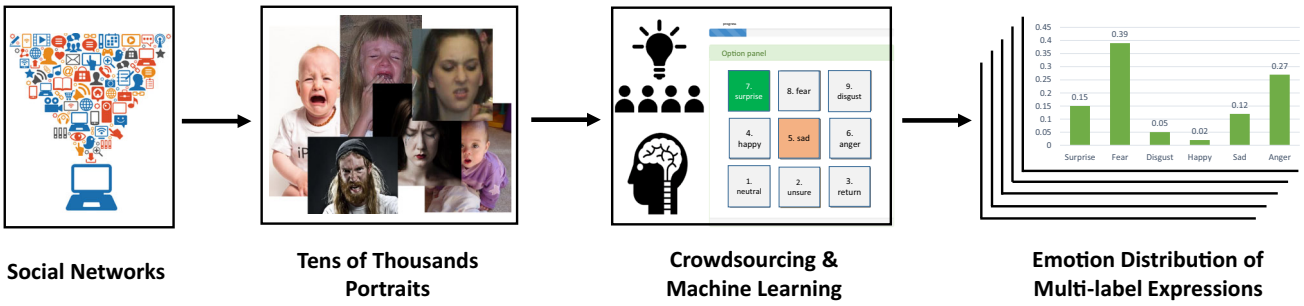


Fig. 2 Overview of construction and annotation of RAF-ML. Tens of thousands portraits were first collected from social networks. The a website was developed to make it easy for our annotators to contribute, by which images can be randomly and equally assigned to each labeler. After label reliability selection, a 6-dimensional distribution ground truth is attached to each sample

(EM) framework, to iteratively optimize and assess the target parameters of each labeler's reliability.

Let $\mathcal{D} = \{(x_j, y_j, t_j^1, t_j^2, \ldots, t_j^R)\}_{j=1}^n$ denote a set of n labeled inputs, where y_j is the gold standard label (hidden variable) for the jth samples $x_j, t_j^i \in \{1, 2, 3, 4, 5, 6, 7\}$ is the corresponding label given by the ith annotator. The correct probability of t_j^i are formulated as a sigmoid function: $p(t_j^i = y_j|\alpha_i, \beta_j) = (1 + \exp(-\alpha_i\beta_j))^{-1} = \sigma(\alpha_i, \beta)$, where $1/\beta_j$ is the difficulty of the jth image, α_i is the reliability of ith annotator.

Our goal is to optimize the log-likelihood of the given labels:

$$\begin{aligned}\max_{\boldsymbol{\beta}>0} l(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \sum_j \ln p(\mathbf{t}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_j \ln \sum_{\mathbf{y}} p(\mathbf{t}, \mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \sum_j \ln \sum_{\mathbf{y}} Q(\mathbf{y}) \frac{p(\mathbf{t}, \mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{Q(\mathbf{y})} \\ &\geq \sum_j \sum_{\mathbf{y}} Q(\mathbf{y}) \ln \frac{p(\mathbf{t}, \mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{Q(\mathbf{y})},\end{aligned}$$

where the last step is deduced by the Jensen's inequality. Instead of explicitly maximize $l(\boldsymbol{\alpha}, \boldsymbol{\beta})$, we choose to iteratively calculate the lower-bound on $l(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in E-step, and then optimize the lower-bound in M-step.

Let $Q(\mathbf{y})$ be a certain distribution of hidden variable $\mathbf{y}$ (i.e., $\sum_{\mathbf{y}} Q(\mathbf{y}) = 1$). According to the necessary and sufficient condition about the equality for the Jensen's inequality, $\frac{p(\mathbf{t}, \mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{Q(\mathbf{y})} = c$ for some constant c that does not depend on $\mathbf{y}$. Then, we can further get:

$$\begin{aligned}Q(y_j) &= \frac{p(\mathbf{t}_j, y_j|\boldsymbol{\alpha}, \beta_j)}{\sum_{\mathbf{y}} p(\mathbf{t}_j, y_j|\boldsymbol{\alpha}, \beta_j)} = \frac{p(\mathbf{t}_j, y_j|\boldsymbol{\alpha}, \beta_j)}{p(\mathbf{t}_j|\boldsymbol{\alpha}, \beta_j)} \\ &= p(y_j|\mathbf{t}_j, \boldsymbol{\alpha}, \beta_j),\end{aligned}$$

here $\mathbf{t}_j$ indicates the set of all given labels for the jth sample. Thus, we set $Q(\mathbf{y})$ to be the posterior distribution of $\mathbf{y}$ given $\mathbf{t}$ and the currently estimated parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

After obtaining the lower-bound on the log-likelihood in E-step, we optimize the lower-bound with respect to the model parameters using gradient ascent method to obtain a new setting of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in M-step. These two steps are iteratively carried out until convergence. Specifically, the joint probabilities in the lower bound can be formulated as:

$$\begin{aligned}P(\mathbf{t}, \mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \prod_j p(y_j) \prod_i p(t_j^i|y_j, \alpha_i, \beta_j) \\ &= \prod_j p(y_j) \prod_i \left(\sum_c p(t_j^i|y_j = c, \alpha_i, \beta_j) \right) \\ &= \prod_j p(y_j) \prod_i \left(\sum_c \left(\sigma(\alpha_i\beta_j)^{\boldsymbol{I}(t_j^i = c)} \right.\right. \\ &\quad \left.\left. * \left(\frac{1}{c-1} \left(1 - \sigma(\alpha_i\beta_j)\right) \right)^{1 - \boldsymbol{I}(t_j^i = c)} \right) \right),\end{aligned}$$

where $\boldsymbol{I}(A)$ is an indicator function that evaluates to "1" if the Boolean expression A is true and "0" otherwise.

After revision, 285 annotators' labels have been remained and Cronbach's Alpha score of all labels is 0.966. Algorithm 1 summarizes the learning process of label reliability estimation. In contrast to the Gaussian prior initialization in Whitehill et al. (2009), we leverage the prior knowledge of annotation for initialization, which shows faster convergence.

3.3 Multi-label Expression Selection

Let $\mathbf{G}_j = (g_1, g_2, \ldots, g_7)$ denote the ground truth probability vector of the jth image, where $g_k = \sum_{i=1}^R \alpha_i \boldsymbol{I}(t_j^i = k)$, and label $k \in \{1, 2, 3, 4, 5, 6, 7\}$ refers to surprise, fear, disgust, happiness, sadness, anger and neutral, respectively. We first calculated the mean distribution value $g_{mean} = \sum_{k=1}^7 g_k/7$ for each image, then chose label k w.r.t. $g_k > g_{mean}$ as the valid label. Images which only have single valid label were filtered out, and the remained 4908 images constitute RAF-ML. This selection criterion, to a large extent, ensures samples in RAF-ML all present multi-label

Algorithm 1 Label reliability estimation algorithm

Input: Training set $\mathcal{D} = \{(x_j, t_j^1, t_j^2, ..., t_j^R)\}_{j=1}^n$

Output: Each annotator's reliability α_i^*

Initialize:

$\forall j = 1, ..., n$, initialize the true label y_j using majority voting. The initial value of β_j is image j's entropy. The higher the entropy, the more uncertain the image.

$$\beta_j := -\sum_{i=1}^{R} p(t_j^i) \ln p(t_j^i) \quad \alpha_i := 1$$

Repeat:

E-step:

$$Q(y_j) := \prod_i p(y_j | t_j^i, \alpha_i, \beta_j)$$

M-step:

$$\alpha_i := \arg\max_{\alpha_i} \sum_j \sum_{\mathbf{y}} Q(y_j) \ln \frac{p(t_j, y_j | \alpha_i, \beta_j)}{Q(y_j)}$$

*We also optimize β_j along with α_i during M-step. However, the goal is to get each labeler's reliability, so we didn't include it in this step. For optimization, we take a derivative with respect to β_j and α_i respectively.

Until convergence

expressions. We further discarded the 7th element (representing 'neutral') and re-normalized, resulting in a new 6-dimensional distribution. We then set the threshold to 1/6 and got the multi-label set for each images. As a result, the emotion category of each image in our RAF-ML is presented as a combination of several basic emotions.

3.4 Dataset Metadata

In RAF-ML, the number of images with two, three and four labels are 3954, 913 and 41, respectively.[4] Specifically, the number of images including each basic emotion are 2093 for Surprise, 1197 for Fear, 2607 for Disgust, 1330 for Happiness, 1647 for Sadness, and 1937 for Anger, respectively. To compare our dataset with the lab-controlled one, some examples of multi-label images and their emotion distributions from RAF-ML and JAFFE are shown in Fig. 3.

We have also provided both manual and automatic modes of facial landmarks. Rough locations and points were first detected automatically using the Viola–Jones face detector (Viola and Jones 2001) and SDM (Xiong and De la Torre 2013). Then, the imprecise or missed detections were corrected by our labelers to get accurate five landmarks (the centers of two eyes, the tip of the nose and the two corners of the mouth). Besides, an automatic landmark annotation mode is included: 37 landmarks were picked out from Face++ API (Inc. 2013). After getting the landmarks, we applied an affine transformation that maps the two eye centers and the center of mouth corners to the fixed coordinates (20, 30), (80, 30) and (50, 75) in 100×100 size, and then gray-scaled the cropped images.

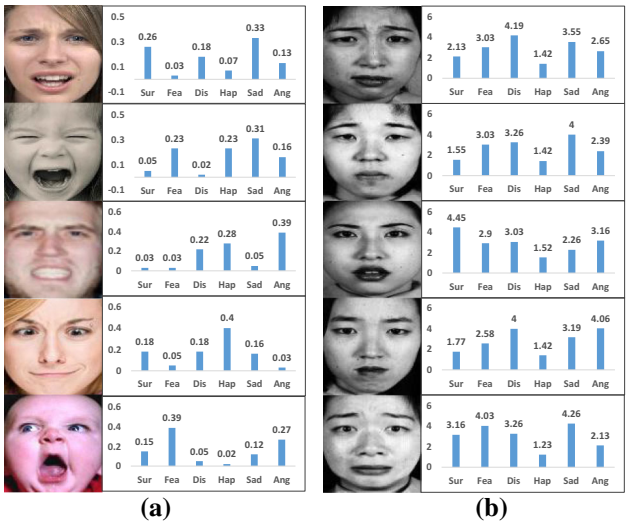


Fig. 3 Example images (already aligned) with their ground truths in RAF-ML and JAFFE. Comparing the emotion distribution histogram of these two databases, we can see that with the crowd-sourced label annotation and the EM optimized algorithm, images in RAF-ML have more distinct multi-label expression distribution. **a** RAF-ML, **b** JAFFE

Beside the facial landmarks, we have also manually annotated other basic attributes (age, gender and race) for RAF-ML. In particular, 9.09, 15.94, 55.47, 15.83 and 3.67 percent of the subjects are in age ranges [0, 3), (4, 19), (20, 39), (40, 69) and (70, –), respectively. For gender distribution, there are 53% female, 45% male, and 2% remains unsure. For racial distribution, there are 74% Caucasian, 10% African American, and 16% Asian.

3.5 Single-Label Classification on RAF-ML

To see if methods that behave well in single-label classification tasks can also gain high recognition rate when tested on multi-label data, we assume RAF-ML to be a basic expression dataset where the predominant emotion of each image is re-selected as the single label ground truth, and then apply it to the single-label classification task.

During experiment, we first extracted deep features from AlexNet (Krizhevsky et al. 2012) and VGG network (Simonyan and Zisserman 2014). For classifier, support vector machine (SVM) with linear kernel implemented by LibSVM (Chang and Lin 2011) was employed, and the penalty parameter C was determined by means of a fivefold cross-validation method. The best single-label classification performance is 42.87% for AlexNet and 43.48% for VGG network. Concrete result for deep features from VGG network is shown in Table 1, in the form of a confusion matrix.

[4] It's reasonable that there are no images with five or six labels in RAF-ML, since people can hardly perceive both negative and positive valence-level (for example, joyful and angry) simultaneously from still images, which occupy different regions in the valence-activation space (Cowie et al. 2001; Russell and Barrett 1999).

Table 1 Confusion matrix for the six basic emotion categories when using VGG features and SVM

(%)	Sur	Fea	Dis	*Hap*	Sad	Ang
Sur	**65.67**	3.86	18.88	3.43	4.29	3.86
Fea	29.23	**23.08**	26.15	3.08	6.15	12.31
Dis	10.41	0	**66.52**	2.26	6.79	14.03
Hap	39.22	1.96	30.39	**6.86**	4.90	16.67
Sad	13.45	0.58	54.39	3.51	**15.20**	12.87
Ang	13.68	2.11	32.11	4.21	6.84	**41.05**

Bold values indicate the results of the diagonal of the confusion matrix

From the matrix, we can see that there exists a very significant degree of confusion between these six basic emotions, which suggests that multi-label emotions analysis in real world can not be simply treated as single-label problem and techniques which can better comprehend the continuum of human emotional behavior should be developed to deal with this special data group.

4 Deep Bi-Manifold (DBM) Feature Learning

Deep convolutional neural networks (DCNNs) have achieved state-of-the-art performance on a wide range of tasks in computer vision community. Furthermore, recent studies found that the softmax loss layer of CNN only encourages the separability of features, which makes Deep Convolutional activation features (DeCaf) (Donahue et al. 2014) not discriminative sufficiently. Several works have appended additional tasks (layers), such as contrastive loss (Hadsell et al. 2006), triplet loss (Schroff et al. 2015) and center loss (Wen et al. 2016), to enhance the discriminative power of the deeply learned features. Unfortunately, these discriminative CNN methods are all designed for single-label recognition tasks, which may not be suitable for our multi-label expression recognition. To address this limitation, we propose a novel deep learning model, called DBM-CNN, to make sense of label distribution information arising from blended expressions, by making two major improvements on the conventional CNN.

Let $\mathcal{D} = \{(x_i, \mathbf{y}_{i\cdot})\}_{i=1}^{n}$ be our dataset with $x_i \in \mathbb{R}$ the ith image and $\mathbf{y}_{i\cdot}$ the corresponding label set. The binary vector $\mathbf{y}_{i\cdot} = (y_i^1, \ldots, y_i^C) \in \{0, 1\}^C$, where C is the total number of classes, and $y_i^j = 1$ when x_i is assigned to the jth expression class and 0 otherwise. Our goal is to learn a multi-label expression model that can capture label distribution information arising from blended expressions and find the optimal bipartite partition of relevant and irrelevant labels.

Firstly, we replace the traditional softmax loss with the multi-label cross-entropy loss which can be defined as:

$$L_c = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C} y_i^j \log p_i^j, \tag{1}$$

where $p_i^j = \frac{exp(f_j(x_i))}{\sum_k exp(f_k(x_i))}$ is the predicted probability of the jth emotion category for the ith input feature output by the softmax, and $f_j(x_i)$ is the jth dimension of the intermediate feature for the ith input sample learned from the network. This cross-entropy loss merely helps to keep the deep features with distant label vectors separable, which is the base of DBM feature learning.

According to the chain rule, the gradient of L_c with respect to $f_j(x_i)$ is computed as:

$$\begin{aligned}\frac{\partial L_c}{\partial f_j(x_i)} &= \sum_{m=1}^{C}\frac{\partial L_c}{\partial p_i^m}\frac{\partial p_i^m}{\partial f_j(x_i)} \\ &= -\frac{1}{n}\sum_{m=1}^{C}\frac{y_i^m}{p_i^m} * p_i^m\left(\boldsymbol{I}(m=j) - p_i^j\right) \\ &= -\frac{1}{n}(y_i^j - p_i^j)\end{aligned} \tag{2}$$

Secondly, and more importantly, we add a new supervised layer on the fundamental architecture shown in Fig. 4, called *bi-manifold loss*, to further enhance the discrimination ability of the deep features. The "bi-manifold" here indicates the layer takes account of the information from both the feature manifold $\mathcal{M}^f$ and the label manifold $\mathcal{M}^l$.

Let the deep feature $x^f \in \mathbb{R}^d$ denote the Deep Convolutional activation feature (DeCaf) of the training sample x, which is assumed to reside on $\mathcal{M}^f$, and the label distribution vector $x^l \in \mathbb{R}^C$ denote the intensities of emotions attached to the relevant sample x, which is assumed to reside on the label manifold $\mathcal{M}^l$. Our objective is to explicitly learn the multi-label classification oriented feature by *aligning the locality of the deep features x^f and corresponding label distribution vector x^l*. Inspired by He and Niyogi (2004), we realize the goal by constructing the nearest-neighbor graph of expression manifold. To formulate the optimization procedure:

$$\min \sum_{i,j}\left(S_{ij}^f + S_{ij}^l\right)\left\|x_i^f - x_j^f\right\|_2^2, \tag{3}$$

where the affinity matrices S^f and S^l specify the similarity between x_i and x_j on the feature manifold and label manifold, respectively. A possible way of defining S^f and S^l is as follows,

$$S_{ij}^f = \begin{cases} 1, & x_j^f \text{ is } k\text{-NN of } x_i^f, \text{ vice versa} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$S_{ij}^l = \begin{cases} 1, & x_j^l \text{ is } k\text{-NN of } x_i^l, \text{ vice versa} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

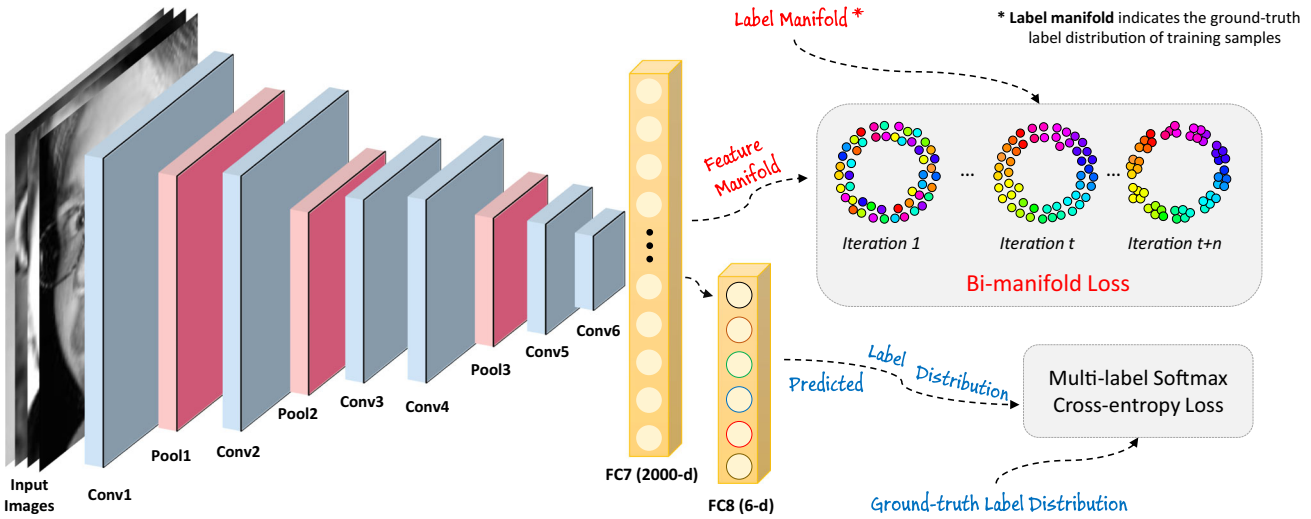


Fig. 4 Framework of the proposed DBM-CNN model. The multi-label softmax cross-entropy loss provides the most basic classification information and keeps features with distant label vectors apart. The bi-manifold loss forces the neighboring features to share similar label vectors and thus become more compact by iteratively aligning the feature manifold and the label manifold

where the local neighborhood is defined by the *k nearest neighbors of both feature* x^f *and label* x^l; that is, we calculate the k-nearest neighbors on the feature manifold $\mathcal{M}^f$ and the label manifold $\mathcal{M}^l$ in the meanwhile and then combine them together. Note that we have also attempted to define the neighborhood by the KNN of the concatenated vector of x^l and x^f, which shows less meanings for the manifold feature learning.

During the training process, the feature space x^f should be updated as the iterative optimization of the CNN. To compute the summation of the pairwise distance, we need to take the entire training set in each iteration, which is inefficient to implement. To address this difficulty, we do the approximation by searching the k nearest bi-neighbors for each sample x_i. Therefore, the bi-manifold loss function is defined as:

$$L_{bm} = \frac{1}{2n}\sum_{i=1}^{n}\left\| 2x_i^f - \frac{1}{k}\sum_{x\in N_k^f\{x_i\}} x^f - \frac{1}{k}\sum_{x\in N_k^l\{x_i\}} x^f \right\|_2^2, \tag{6}$$

where $N_k^f\{x_i\}$ and $N_k^l\{x_i\}$ denotes the ensemble of the k nearest neighbors of sample x_i on the feature manifold and the label manifold respectively. The gradient of L_{bm} with respect to x_i^f is computed as:

$$\frac{\partial L_{bm}}{\partial x_i^f} = \frac{1}{n}\left(2x_i^f - \frac{1}{k}\sum_{x\in N_k^f\{x_i\}} x^f - \frac{1}{k}\sum_{x\in N_k^l\{x_i\}} x^f \right). \tag{7}$$

In the initial period of the training, there is no obvious overlap between $\mathcal{M}^f$ and $\mathcal{M}^l$ due to unconstrained variations such as the illuminations, poses and individual appearances. As the learning process goes on, the coincidence degree of the k-nearest neighbors on x^f and x^l will increase, and the joint alignment of the feature manifold and the label manifold will also be realized. In other word, the two manifolds would *progressively align* in such a way that the neighboring samples tend to share similar multi-label distribution vector.

In summary, we adopt the joint supervision of cross-entropy loss, which characterizes the global scatter, and the bi-manifold loss, which characterizes the local scatters within similar intensity distribution information, to train the CNNs for discriminative feature learning. The objective function is formulated as follow:

$$L = L_c + \lambda L_{bm}, \tag{8}$$

where L_c denotes the cross-entropy loss and L_{bm} denotes the bi-manifold loss. The hyper parameter λ is used to balance these two loss functions. In this manner, we can perform the update based on mini-batch. Algorithm 2 summarizes the learning process in the DBM-CNN. Intuitively, the cross-entropy loss forces the deep features with different labels to stay apart and the bi-manifold loss efficiently pulls the neighboring deep features with the similar emotion distribution together. With the joint supervision, both the inter-class feature differences and the local feature correlations are enlarged. Hence the discriminative power of the deeply learned features can be highly enhanced.

Algorithm 2 Optimization algorithm of DBM-CNN.

Input: The training data $\mathcal{D} = \{(x_i, \mathbf{y}_i.)\}_{i=1}^{n}$, n is the size of mini-batch.

Output: Network layer parameters Θ.

Initialize: The number of iteration $t \leftarrow 0$, network learning rate μ, hyper parameter λ, network layer parameters Θ, bi-neighboring nodes k.

Repeat:

1: $t \leftarrow t + 1$

2: Computer each center of k-nearest bi-neighbor for x_i:
$C_i^t = \frac{1}{k} \sum_{i \neq j}^{n} x_j^{f^{(t)}} \left(S_{ij}^{f^{(t)}} + S_{ij}^{l} \right)$

3: Compute the joint loss:
$L^t = L_c^t + \lambda L_{bm}^t$

4: Update the back-propagation error for x_i by Eq (2) and Eq (7):
$\frac{\partial L^t}{\partial x_i^{f^{(t)}}} = \frac{\partial L_c^t}{\partial x_i^{f^{(t)}}} + \lambda \frac{\partial L_{bm}^t}{\partial x_i^{f^{(t)}}}$

5: Computer the network layer parameters Θ:
$\Theta^{t+1} = \Theta^t - \mu^t \frac{\partial L^t}{\partial \Theta^t} = \Theta^t - \mu^t \sum_{i=1}^{n} \frac{\partial L^t}{\partial x_i^{f^{(t)}}} \frac{\partial x_i^{f^{(t)}}}{\partial \Theta^t}$

Until convergence

5 Domain Adaption for Extensive Facial Expression Recognition Tasks

Due to differences in collection conditions, subject characteristics and behaviors, facial expression inconsistency is common across databases and cultures (Jack et al. 2012), which makes it challenging to conduct expression recognition in the cross-dataset scenario. Therefore, in this section, deep bi-manifold features learned from RAF-ML is further exploited via domain adaptation so that it can generalize well to other related databases for a wide range of emotion analysis tasks.

Domain adaptation (DA) has been widely utilized in cross-database scenarios where the training data in source domains used to learn a model has different distribution from the target data (Patel et al. 2015; Csurka 2017). Traditional domain adaptation methods assume that the task is the same, i.e., class labels are shared between domains. Therefore, the feature learned from RAF-ML can appropriately transfer to other databases within multi-label expression recognition task. In addition, since the multi-label emotions are composed of two or more component categories of basic expression, and our DBM-CNN learns features for multi-label facial expressions by exploring the emotion intensity distribution information from these basic components, we believe that the DBM-CNN architecture can also learn competent features for basic expression recognition by the agency of domain adaption technique.

Considering that in this context training and test data are drawn from different distribution, Maximum Mean Discrepancy (MMD) which is an effective metric for comparing the distributions is employed to decrease the cross-domain discrepancy in feature space. Denote by the source domain $\mathcal{S} = \{(x_i^s, \mathbf{y}_i.)\}_{i=1}^{n_s}$ the training multi-label data and the target domain $\mathcal{T} = \{x_i^t\}_{i=1}^{n_t}$ the test data without label information, where $\mathbf{y}_i.$ denotes the label vector of the sample x_i^s, and $x_i^s, x_i^t \in \mathbb{R}^d$. The MMD between these two domains and its empirical estimate in the reproducing kernel Hilbert space (RKHS) can be defined as:

$$\mathrm{MMD}[\mathcal{F}, p_s, p_t] := \sup_{f \in \mathcal{F}} \left(\mathbf{E}_{p_s}[f(\mathbf{x}^s)] - \mathbf{E}_{p_t}[f(\mathbf{x}^t)] \right), \quad (9)$$

$$\mathrm{MMD}[\mathcal{H}, \mathcal{S}, \mathcal{T}] := \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(x_i^s) - \frac{1}{n_t} \sum_{i=1}^{n_t} \phi(x_i^t) \right\|_{\mathcal{H}}, \quad (10)$$

where $\mathbf{E}_{p_s}$ and $\mathbf{E}_{p_t}$ denote the population expectations under distribution p_s and p_t, respectively. The MMD function class $\mathcal{F}$ is the unit ball in a reproducing kernel Hilbert space $\mathcal{H}$. According to Gretton et al. (2012a), $\mathrm{MMD}[\mathcal{F}, p_s, p_t] = 0$ if and only if $p_s = p_t$. Hence we can use the MMD to detect any discrepancy between p_s and p_t.

By mapping the data into $\mathcal{H}$ using feature space mapping function $\phi(\cdot)$, we can get the kernel map $k(x^s, x^t) = \langle \phi(x^s), \phi(x^t) \rangle_{\mathcal{H}}$. One of the most used kernel corresponding to an infinite-dimensional $\mathcal{H}$ is the Gaussian kernel $k(x^s, x^t) = exp(-||x^s - x^t||^2/(2\sigma^2))$. To maximize the test power, Gretton et al. (2012b) proposed a new kernel selection approach and denoted a linear combination of base kernels $\{k_l\}_{l=1}^{d}$:

$$\mathcal{K} := \left\{ k = \sum_{l=1}^{d} \beta_l k_l, \sum_{l=1}^{d} \beta_l = 1, \beta_l \geqslant 0, \forall l \right\}. \quad (11)$$

Under the above assumptions, Gretton et al. (2012a) further proposed an unbiased estimator of $\mathrm{MMD}^2[\mathcal{H}, \mathcal{S}, \mathcal{T}]$:

$$\begin{aligned} \mathrm{MMD}_u^2[\mathcal{H}, \mathcal{S}, \mathcal{T}] = & \frac{1}{n_s(n_s-1)} \sum_{i \neq j}^{n_s} k(x_i^s, x_j^s) \\ & + \frac{1}{n_t(n_t-1)} \sum_{i \neq j}^{n_t} k(x_i^t, x_j^t) \\ & - \frac{2}{n_s n_t} \sum_{i,j=1}^{n_s, n_t} k(x_i^s, x_j^t). \end{aligned} \quad (12)$$

Following the work in Long et al. (2015), a multi-kernel MMD loss layer is appended to DBM-CNN. As related

research has suggested (Yosinski et al. 2014), transferability of the network representation sharply decreases as the layer goes deeper and hence it will become difficult to directly transfer the learned feature to the target domain. Therefore, we utilize the output of the feature embedding layer as the RKHS for the MK-MMD loss layer so as to regularize the learned representation to be invariable to domain shift.

By adding the MK-MMD loss layer into the network, the objective function can be formulated as: $L = L_c + \lambda L_{bm} + \gamma L_{mmd}$, where L_c and L_{bm} denote the cross-entropy loss and the bi-manifold loss on the source domain respectively, and L_{mmd} is the multi-kernel MMD loss. λ and γ are the hyper-parameters to weight against these loss functions. With the definition in Eq. (12), the gradients of the MMD loss with respect to source feature x_i^s and target feature x_i^t can be computed as:

$$\begin{aligned}\frac{\partial L_{mmd}}{\partial x_i^s} &= \frac{1}{n_s(n_s-1)}\sum_{i\neq j}^{n_s}\frac{\partial k(x_i^s, x_j^s)}{\partial x_i^s} \\ &\quad - \frac{2}{n_s n_t}\sum_{i,j=1}^{n_s,n_t}\frac{\partial k(x_i^s, x_j^t)}{\partial x_i^s},\end{aligned} \tag{13}$$

$$\begin{aligned}\frac{\partial L_{mmd}}{\partial x_i^t} &= \frac{1}{n_t(n_t-1)}\sum_{i\neq j}^{n_t}\frac{\partial k(x_i^t, x_j^t)}{\partial x_i^t} \\ &\quad - \frac{2}{n_s n_t}\sum_{j,i=1}^{n_s,n_t}\frac{\partial k(x_j^s, x_i^t)}{\partial x_i^t}.\end{aligned} \tag{14}$$

Given the Gaussian multi-kernel defined in Eq. (11), we typically take $\partial k(x_i^s, x_j^t)/x_i^s$ for example:

$$\frac{\partial k(x_i^s, x_j^t)}{\partial x_i^s} = -\sum_{l=1}^{d}\frac{\beta_l}{\sigma_l^2}k_l(x_i^s, x_j^t) * (x_i^s - x_j^t). \tag{15}$$

In addition, as the domain adaption setting is unsupervised that only labels in the source domain are provided and labels in the target domain remain unknown, the learned label space on the target domain will tend to present multimodal distribution given the supervision information from the multi-label source domain, which is contrary to the actual single label it holds. So, an entropy loss layer is further appended to the last fully-connected layer of the network for this special scenario. The entropy loss layer is formulated as follow:

$$L_e = -\frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{j=1}^{C} p_i^j \log(p_i^j), \tag{16}$$

where $p_i^j = \frac{exp(f_j(x_i^t))}{\sum_k exp(f_k(x_i^t))}$ is the predicted probability of the jth emotion category for the ith input target feature and $f_j(x_i^t)$ is the jth dimension of the intermediate feature x_i^t

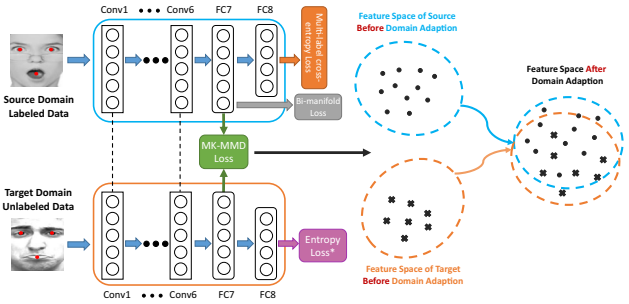


Fig. 5 Framework of the domain adaption network DBM-DACNN. Input images from the source domain and the target domain are aligned to the uniform template and share the same fundamental architecture (from $Conv1$ to $Fc8$). Specifically, the entropy loss works only when the target domain is single-label data

learned from the network. And C is the total number of classes. The entropy loss reaches the maximum when the probabilities are uniform distribution. And it reaches zero only if one entry of the output probabilities is 1 and all others are 0.

According to the chain rule, the gradient of L_e with respect to $f_j(x_i^t)$ is:

$$\begin{aligned}\frac{\partial L_e}{\partial f_j(x_i^t)} &= \sum_{m=1}^{C}\frac{\partial L_e}{\partial p_i^m}\frac{\partial p_i^m}{\partial f_j(x_i^t)} \\ &= -\frac{1}{n_t}\sum_{m=1}^{C}\left(\log(p_i^m)+1\right) * p_i^m\left(\boldsymbol{I}(m=j) - p_i^j\right) \\ &= -\frac{1}{n_t}p_i^j\left(\sum_{m=1}^{C} p_i^m \log(p_i^m) - \log(p_i^j)\right).\end{aligned} \tag{17}$$

By updating the network parameters with mini-batch SGD, the cross-domain network DBM-DACNN can dig into the intensity differences among basic component expressions on the RAF-ML, and in the meantime, thoroughly adapt this important information to the target data by bridging the domain discrepancy. Moreover, by minimizing the label entropy on the target single-label data, the learned label space can be forced to exhibit an unimodal distribution that is closer to the ground truth. Thus, the discriminative power of the learned features can generalize well to extensive facial expression recognition tasks on other related expression datasets. Figure 5 and Algorithm 3 shows the framework and the learning process of DBM-DACNN, respectively.

6 Experiments

In this section, we first evaluate the effectiveness of the DBM-CNN features on RAF-ML. Then the extended domain

Algorithm 3 Optimization algorithm of the DBM-DACNN.

Input: Source data $\mathcal{S} = \{(x_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$,
Target data $\mathcal{T} = \{x_i^t\}_{i=1}^{n_t}$,
n_s and n_t is the size of mini-batch.

Output: Network layer parameters Θ.

Initialize: The number of iteration $j \leftarrow 0$, network learning rate μ, bi-neighboring nodes k, hyper parameter λ and γ, network layer parameters Θ.

Repeat:

1: $j \leftarrow j + 1$

2: Compute the joint loss:
M2M†: $L^j = L_c^j + \lambda L_{bm}^j + \gamma L_{mmd}^j$
M2S‡: $L^j = L_c^j + \lambda L_{bm}^j + \gamma L_{mmd}^j + \eta L_e$

3: Compute the back-propagation error for x_i by Eq. (13), Eq. (14) and Eq. (17):

$$\frac{\partial L^j}{\partial x_i^{s(j)}} = \frac{\partial L_c^j}{\partial x_i^{s(j)}} + \lambda \frac{\partial L_{bm}^j}{\partial x_i^{s(j)}} + \gamma \frac{\partial L_{mmd}^j}{\partial x_i^{s(j)}}$$

$$\frac{\partial L^j}{\partial x_i^{t(j)}} = \gamma \frac{\partial L_{mmd}^j}{\partial x_i^{t(j)}} + \eta \frac{\partial L_e}{\partial x_i^t}$$

4: Update the network layer parameters Θ:

$$\Theta^{j+1} = \Theta^j - \mu^j \frac{\partial L^j}{\partial \Theta^j}$$
$$= \Theta^j - \mu^j \left(\sum_{i=1}^{n_s} \frac{\partial L^j}{\partial x_i^{s(j)}} \frac{\partial x_i^{s(j)}}{\partial \Theta^j} + \sum_{i=1}^{n_t} \frac{\partial L^j}{\partial x_i^{t(j)}} \frac{\partial x_i^{t(j)}}{\partial \Theta^j}\right)$$

Until convergence

M2M†: Multi-label source domain to Multi-label target domain
M2S‡: Multi-label source domain to Single target domain

adaption network DBM-DACNN are employed to adapt features learned on RAF-ML to other related expression databases: multi-label expression database JAFFE and three commonly-used single-label facial expression databases, namely CK+, SFEW 2.0 and MMI. As our methods are generic, we further conduct extensive experiments on other large-scale tasks in the Supplementary Material.

6.1 Data Pre-processing

We tested the algorithms on five diverse datasets that vary in annotation, collection environment, image quality and property of expressions (posed or spontaneous).

(1) RAF-ML is a real-world database which contains 4908 great-diverse facial images downloaded from the Internet with manually annotated multiple labels. During experiment, the dataset has been divided into a training set of 4090 images and a test set of 818 images to ensure that results can be accurately reproduced.
(2) JAFFE (Lyons et al. 1998) is a lab-controlled database which only contains 213 samples from 10 Japanese females with posed expressions. Emotion intensity distribution information for each sample have been provided in this dataset. Images with emotional expressions have been used in our experiments.
(3) CK+ (Lucey et al. 2010) is a lab-controlled database with 593 video sequences from 123 subjects across different culture. Only 309 sequences have been labeled with six basic expression labels. We then extracted the final frame of each sequences with peak formation, resulting in 309 images.
(4) MMI (Valstar and Pantic 2010) is a lab-controlled database which includes 30 subjects with non-uniformly posed expressions and various accessories. We selected the three peak frames in each sequence as basic expressions, resulting in 528 images.
(5) SFEW 2.0 (Dhall et al. 2015b) is a real-world database that contains images selected from different films with spontaneous expressions, various head pose, age range, occlusions and illuminations. The database is divided into three sets for training, validation and testing. And we used images in the training and validation parts that are provided with expression labels.

All the facial images were first aligned to uniform template using two eye centers and the mouth center, then cropped to the 100×100 size and transformed to gray scale for the following feature extraction and classification.

6.2 Architecture and Training Details

All of our models were trained base on an open source deep learning framework, Caffe library (Jia et al. 2014). The already aligned gray-scale images were firstly normalized through dividing all the pixel value by 255. We then considered a network taking a fixed-size input (90×90) cropped from images, which was for the purpose of data augmentation.

In order to compare different models fairly, we adopted uniform training methods and trained the uniform fundamental network architectures shown in Table 2 from scratch. The dropout layer was applied to the last fully connected layers with rate 0.3. The learning rate was initially set to 0.01 and decreased by factor of 10 at 5k and 18k iterations, and we stopped training at 20k iterations. Moreover, we chose the stochastic gradient descent (SGD) as optimization and used mini-batch with 64 samples. The momentum coefficient was set to 0.9.

The model was regularized using weight decay. We set the weight decay coefficient of convolution layer and first fully connect layer to 0.0005 while the second fully connect layer to 0.0025. MSRA was used to initialize the weight parameter of convolutional layer and fully connect layer, while the bias parameter was set to 0 at the beginning of training. All of

Table 2 The configuration parameters in the fundamental architecture (baseDCNN)

Layer type	1 Conv	2 ReLu	3 MPool	4 Conv	5 ReLu	6 MPool	7 Conv	8 ReLu	9 Conv	10 ReLu	11 MPool	12 Conv	13 ReLu	14 Conv	15 ReLu	16 FC	17 ReLu	18 FC
Kernel	3	–	2	3	–	2	3	–	3	–	2	3	–	3	–		–	
Output	64	–	–	96	–	–	128	–	128	–	–	256	–	256	–	2000	–	7
Stride	1	1	2	1	1	2	1	1	1	1	2	1	1	1	1		1	
Pad	1	0	0	1	0	0	1	0	1	0	0	1	0	1	0		0	

our models were trained on a NVIDIA Tesla K40 GPU and it cost about 2 hours to train a model.

6.3 Multi-label Facial Expression Recognition on RAF-ML

To study the multi-label expression in-the-wild problem, we first conducted the multi-label classification experiments on our RAF-ML database.

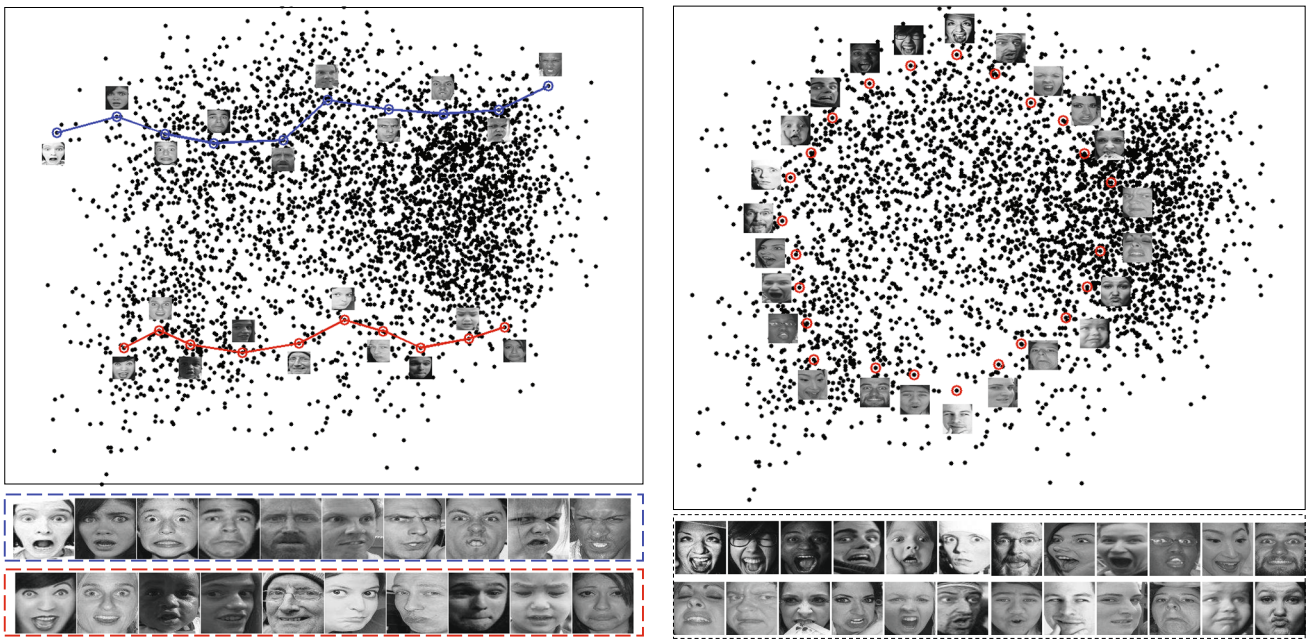
6.3.1 Comparison Features

For the comparison purpose, we implemented two handcrafted and four deeply learned features. For handcrafted features, we employed Local binary patterns (LBP) (Ojala et al. 2002) and histogram of orientated gradients (HOG) (Dalal and Triggs 2005). LBP descriptor applied the 59-bin $LBP^{u2}_{8,2}$ operator and concatenated histograms from 10×10 pixel cells, generating a 5900 dimensional feature vector. HOG feature used the shape-based segmentation dividing the image into 10×10 pixel blocks of four 5×5 pixel cells with no overlapping. By setting 10 bins for each histograms, we extracted a 4000-dimensional feature vector for each image.

To obtaining competitive baseline of the deeply learned feature, a baseDCNN of the same architecture with DBM-CNN was first trained on the RAF-ML training set only using cross-entropy loss, and 2000-dimensional deep activation features were then extracted from the penultimate fully connected layer. Previous studies Donahue et al. (2014) and Sharif Razavian et al. (2014) also proved that well-trained deep convolutional network can work as a generic feature extraction tool with generalization ability for various visual recognition tasks. Motivated by this finding, two widely-used pre-trained object recognition models, namely VGG network (Simonyan and Zisserman 2014) and AlexNet (Krizhevsky et al. 2012) were also employed to directly extract features in our experiments. During parameter optimization on DBM-CNN, we conducted fivefold cross-validation on the training set. And the value of k and λ in DBM-CNN were set to be 10 and 0.01, respectively.

6.3.2 Multi-label Classifiers

To fairly compare the effectiveness of these features, several widely used multi-label learning algorithms have been employed: Calibrated Label Ranking (CLR) (Fürnkranz et al. 2008), Random k-Labelsets (RAkEL) (Tsoumakas and Vlahavas 2007), Multi-Label k-Nearest Neighbor (MLkNN) (Zhang and Zhou 2007), Multi-Label learning using LOcal Correlation (ML-LOC) (Huang et al. 2012), multi-label learning with Label specIfic FeaTures (LIFT) (Zhang and Wu 2015) and Multi-Label Manifold Learning (ML^2) (Hou et al. 2016). For CLR and RAkEL, we used J48 (C4.5) as



(a) **(b)**

Fig. 6 Two-dimensional deep feature embedding by DBM-CNN on RAF-ML, where the corresponding facial images continuously and smoothly change in expression intensity, reflecting the intrinsic structure of expression manifold. By jointly preserving local structure and the emotion distribution information of the deep features, the DBM-CNN implicitly emphasizes the natural clusters in the data and preserves the smooth change within cluster

the fundamental classifier. In addition, k is 3 and n is 12 in RAkEL. For MLkNN, the number of nearest neighbors considered k is set to be 100 and the smoothing parameter is set to be 1. For ML-LOC, the length of the loc code m is set to be 15 and RBF kernel is used in the SVM model. For LIFT, the ratio parameter r is set to be 0.1 and RBF kernel is used in the SVM model. For ML^2, the number of neighbors k is set to be 7 and the parameter C_1, C_2 and λ are all set to be 1.

6.3.3 Evaluation Criteria

To report the performance comprehensively, several commonly-used evaluation criteria in multi-label learning were applied. Denoting the number of examples is N, the number of labels is L. $y_{i,j}$ is the ground truth of the ith sample on jth label. $\hat{y}_{i,j}$ is the prediction of the ith sample on jth label. We use $|\cdot|$ to denote the cardinality of a set, i.e., the number of relevant labels in the set. The brief descriptions of the measurements are given in the following:

(1) Hamming Loss: This criterion measures the degree of inconsistency between the predicted results and the ground truth of the sample, i.e., the possibility of a relevant label is missed or an irrelevant is predicted. It is formally defined as a fraction of the wrong labels to the total number of labels:

$$\frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L}\boldsymbol{I}(y_{i,j}\neq\hat{y}_{i,j}).$$

(2) Coverage: This metric measures the number of labels on average that should be included to cover all relevant labels in the ranking queue. It is formally described as:

$$\frac{1}{N}\sum_{i=1}^{N}\max_{j:\,y_{i,j}=1}|\{k:\hat{y}_{i,k}\geq\hat{y}_{i,j}\}|.$$

(3) One-error: This indicator describes the degree of the top-ranked predicted label is not in set of true class labels of the instances. For single-label learning, it means the classification error. Formally, it is represented as:

$$\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{I}\left(\hat{y}_{i,0}\notin\{j:y_{i,j}=1\}\right),$$

where $\hat{y}_{i,0}$ is the top ranked label in $\hat{\mathbf{y}}_i$.

(4) Ranking Loss: This criterion describes the average fraction of label pairs miss-ordered of each instance, that is, the probability that the irrelevant labels are ranked higher than the relevant ones. It is formally described as:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\left|\left\{(j,k):\hat{y}_{i,j}<\hat{y}_{i,k},\, y_{i,j}=1,\, y_{i,k}=0\right\}\right|}{|\mathbf{y}_{i\cdot}|(L-|\mathbf{y}_{i\cdot}|)}.$$

(5) Average precision: This criterion is based on the notion of label ranking and reflects the average fraction of relevant labels ranked higher than one other relevant label. It can be formulated as:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathbf{y}_{i\cdot}|}\sum_{j:\,y_{i,j}=1}\frac{\left|\left\{k:\,y_{i,k}=1,\,\hat{y}_{i,k}\geq\hat{y}_{i,j}\right\}\right|}{|\{k:\,\hat{y}_{i,k}\geq\hat{y}_{i,j}\}|}.$$

(6) Micro- and Macro-F_1: The traditional F-measure (F_1 score) can be interpreted as the weighted harmonic mean value of precision and recall, where precision is the ability of the classifier not to label a sample that is negative as positive and recall is the probability of a classifier to find all the positive samples. For Micro-averaged method, the individual dividends and divisors that make up the per-class metrics are summed up to calculate an overall quotient. And Macro-averaged method simply calculates the average of the binary metrics, giving equal weight to each class.

(7) Micro- and Macro- AUC: AUC used here generically refers to the area under the receiver operating characteristic (ROC) curve. The ROC curve visualizes the trade-off between sensitivity and specificity by plotting both values as a function of a varying classification threshold. And the larger value of AUC is, the better performance of the corresponding classifier is. Micro-AUC calculates AUC on prediction matrix, and Macro-AUC calculates AUC averaging on each label.

6.3.4 Experimental Results

Figure 6 shows the resulting 2-dimensional deep features learned from our DBM-CNN model, where we attach example face images with various expression intensity information. And Table 3 enumerates the comparative multi-label expression recognition performances of six features across different classification algorithms on RAF-ML.

From the comparison results, we make three interesting observations. First, the pre-trained models AlexNet and VGG network which achieve quite reasonable results for large-scale image recognition settings, are not discriminative enough for expression recognition tasks. And they even perform worse than the unlearned handcrafted features in some cases. Second, baseDCNN outperforms both the hand-engineered features and pre-trained models by a significant margin, which indicates that training on RAF-ML with expression label information has learned useful emotion-specific features that are suitable for multi-label expression recognition. Last but not least, the proposed DBM-CNN model achieves the most competitive performance under all test cases that cover various evaluation criteria and different classification methods, which suggests that the bi-manifold loss indeed helps to enhance the discriminative ability of the deep features by jointly preserving the local compactness of both label manifold and feature manifold.

6.4 Discussions on DBM-CNN

In this section, we look deeper into the DBM-CNN and explore different parameter settings and manifold selections to see how these changes influence the performance on DBM-CNN.

6.4.1 Parameter Sensitivity

The objective of DBM-CNN consists of two terms, i.e., cross-entropy empirical loss and bi-manifold loss. Both of this two regulations are of essential importance for multi-label expression recognition. The parameter λ makes a trade off between this two parts, and has a great influence on the performance. The parameter k in bi-manifold loss controls the degree of local concentration in feature and label manifolds. If k is too large, it will urge all samples in the training set to be together; if k is too small, the bi-manifold loss will make no difference.

To investigate the effects of different values of hyper-parameter λ and k used in DBM-CNN model, we conducted two experiments on the multi-label expression recognition task. The performances of these models on RAF-ML are shown in Fig. 7. In the first experiment (left), we fixed k to 10 and varied λ in the set $\{0, 0.001, 0.005, 0.01, 0.05\}$ to learn different models. As the results show, the performances are sensitive to the choice of λ, and $\lambda = 0$ is the case of simply using the cross-entropy loss, which leads to relatively poor performance of the deeply learned features. This confirms the promoting effect of the jointly supervision of cross-entropy loss and bi-manifold loss when a proper trade-off is chosen. In the second experiment (right), we fixed $\lambda = 0.01$ and varied k in the set $\{5, 10, 20, 40\}$ to learn different models. We can observe that the DBM-CNN accuracy first increases and then decreases as k grows and we archive our best performance when k is set to 10.

6.4.2 Manifold Choice

To have a closer look at the effectiveness of the bi-manifold loss that characterizes the feature locality on both feature and label space, we implemented the simplified versions of DBM-CNN with only feature manifold loss or only label manifold loss. Table 4 shows the comparative results on these different schemes using MLKNN classification algorithm. One can observe that both the feature and label manifold loss can help improve the performance of the deep features

Table 3 Experimental results of comparing features on **RAF-ML** using different algorithms

Comparing features	Evaluation criterion								
	Hamming↓	Coverage↓	One-error↓	Ranking↓	Ave. precision↑	Micro-F_1↑	Macro-F_1↑	Micro-AUC↑	Macro-AUC↑
(a) CLR									
LBP	0.423	3.392	0.503	0.404	0.626	0.468	0.449	0.608	0.592
HOG	0.377	2.995	0.394	0.318	0.695	0.537	0.512	0.691	0.667
AlexNet	0.347	2.887	0.349	0.293	0.720	0.548	0.514	0.716	0.684
VGG	0.339	2.780	0.319	0.276	0.737	0.563	0.528	0.735	0.706
baseDCNN	0.279	2.399	0.249	0.205	0.796	0.642	0.616	0.811	0.792
DBM-CNN	**0.217**	**2.133**	**0.159**	**0.154**	**0.848**	**0.716**	**0.690**	**0.860**	**0.846**
(b) RAkEL									
LBP	0.392	3.366	0.513	0.407	0.624	0.444	0.413	0.601	0.578
HOG	0.347	3.119	0.412	0.345	0.677	0.513	0.463	0.668	0.638
AlexNet	0.310	2.860	0.325	0.292	0.728	0.551	0.501	0.722	0.690
VGG	0.308	2.867	0.364	0.301	0.718	0.557	0.503	0.717	0.686
baseDCNN	0.249	2.475	0.230	0.214	0.795	0.657	0.615	0.810	0.785
DBM-CNN	**0.177**	**2.139**	**0.169**	**0.155**	**0.851**	**0.755**	**0.728**	**0.873**	**0.858**
(c) MLkNN									
LBP	0.260	2.564	0.258	0.238	0.775	0.602	0.533	0.789	0.764
HOG	0.251	2.523	0.244	0.228	0.783	0.615	0.563	0.801	0.779
AlexNet	0.278	2.667	0.284	0.256	0.756	0.544	0.454	0.768	0.737
VGG	0.277	2.728	0.271	0.256	0.760	0.546	0.452	0.765	0.736
baseDCNN	0.193	2.105	0.151	0.151	0.854	0.728	0.695	0.873	0.860
DBM-CNN	**0.168**	**1.965**	**0.133**	**0.128**	**0.873**	**0.768**	**0.739**	**0.901**	**0.891**

Table 3 continued

Comparing features	Evaluation criterion								
	Hamming$^{\downarrow}$	Coverage$^{\downarrow}$	One-error$^{\downarrow}$	Ranking$^{\downarrow}$	Ave. precision$^{\uparrow}$	Micro-$F_1^{\uparrow}$	Macro-$F_1^{\uparrow}$	Micro-AUC$^{\uparrow}$	Macro-AUC$^{\uparrow}$
(d) ML-LOC									
LBP	0.259	2.561	0.247	0.232	0.777	0.588	0.524	0.802	0.798
HOG	0.256	2.559	0.246	0.223	0.786	0.612	0.574	0.825	0.802
AlexNet	0.267	2.662	0.285	0.264	0.756	0.584	0.545	0.754	0.731
VGG	0.276	2.660	0.294	0.255	0.759	0.608	0.583	0.767	0.752
baseDCNN	0.189	2.152	0.168	0.156	0.852	0.722	0.693	0.867	0.864
DBM-CNN	**0.173**	**2.079**	**0.139**	**0.135**	**0.864**	**0.744**	**0.722**	**0.886**	**0.876**
(e) LIFT									
LBP	0.251	2.487	0.229	0.218	0.785	0.623	0.542	0.812	0.809
HOG	0.230	2.383	0.211	0.201	0.804	0.653	0.586	0.826	0.815
AlexNet	0.264	2.616	0.253	0.238	0.766	0.614	0.562	0.782	0.769
VGG	0.247	2.540	0.234	0.221	0.782	0.642	0.581	0.813	0.798
baseDCNN	0.187	2.097	0.150	0.147	0.860	0.740	0.715	0.886	0.875
DBM-CNN	**0.167**	**1.954**	**0.119**	**0.116**	**0.879**	**0.769**	**0.745**	**0.903**	**0.894**
(f) ML^2									
LBP	0.254	2.507	0.230	0.222	0.782	0.596	0.526	0.807	0.804
HOG	0.228	2.380	0.206	0.199	0.808	0.656	0.589	0.832	0.822
AlexNet	0.255	2.517	0.247	0.236	0.772	0.630	0.576	0.811	0.789
VGG	0.247	2.537	0.229	0.219	0.785	0.645	0.599	0.825	0.810
baseDCNN	0.187	2.151	0.146	0.153	0.853	0.731	0.698	0.881	0.871
DBM-CNN	**0.171**	**2.025**	**0.128**	**0.133**	**0.869**	**0.757**	**0.731**	**0.897**	**0.886**

For each evaluation criterion, "↓" indicates the smaller the better while "↑" indicates the bigger the better. The best performance on each evaluation criterion is highlighted in bold face

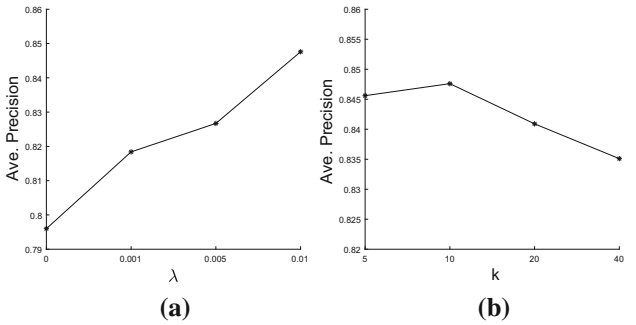


Fig. 7 Multi-label expression recognition performances (Average Precision using CLR classification method) on RAF-ML for different values of λ (left) and k (right). **a** DBM-CNN models with different λ and fixed $k = 10$, **b** DBM-CNN models with different k and fixed $\lambda = 0.01$

individually. Preserving the feature locality can enhance the local clusters, and maintaining the label smoothness can benefit the multi-label classification. By jointly aligning these two manifolds, the proposed bi-manifold CNN outperforms both of them, which indicates that DBM-CNN indeed brings in the superposed advantage that emphasizes the overlap of these two manifolds.

6.5 Domain Adaption to Other Expression Datasets

Conducting expression recognition tasks across datasets is challenging due to variances in acquisition and different settings on people's age range, gender, culture and the level of expressiveness. Figure 8 exhibits example samples from different databases that vary in background, pose, occlusions, illumination and subject identity, which may bias features.

To evaluate the generalization ability of our models to extensive facial expression recognition tasks, we further employed the learned deep manifold features from RAF-ML to other related facial expression databases: JAFFE for multi-label expressions recognition and CK+, MMI and SFEW 2.0 for basic expressions recognition. Faces from the source domain and target domain were first aligned to the uniform template using three reference points (two eye centers and the mouth center).

6.5.1 Multi-label to Multi-label

For multi-label classification task on JAFFE, we applied a threshold of 3 to obtain the label set of each image according to the five-scale (1–5) intensity principle: relevant emotions whose value is greater than 3 are set as 1 and the irrelevant emotions are set as 0. After eliminating images whose intensities for six expressions are all less than 3, we obtained 188 images for the multi-label classification. The same compared features, classification algorithms and evaluation criteria in Sect. 6.3 were used for JAFFE. Considering the only so

Table 4 Comparison results of DBM-CNN and other training models using MLkNN

	Hamming loss$^{\downarrow}$	Coverage$^{\downarrow}$	One-error$^{\downarrow}$	Ranking loss$^{\downarrow}$	Average precision$^{\uparrow}$	Micro-$F_1^{\uparrow}$	Macro-$F_1^{\uparrow}$	Micro-AUC$^{\uparrow}$	Macro-AUC$^{\uparrow}$
baseDCNN	0.193	2.105	0.151	0.151	0.854	0.728	0.695	0.873	0.860
DBM-CNN (S^f)	0.183	2.053	0.146	0.145	0.858	0.746	0.717	0.887	0.875
DBM-CNN (S^l)	0.182	2.061	0.142	0.141	0.863	0.742	0.709	0.885	0.872
DBM-CNN ($S^f + S^l$)	**0.168**	**1.965**	**0.133**	**0.128**	**0.873**	**0.768**	**0.739**	**0.901**	**0.891**

Bold values indicate the best result in term of each performance index

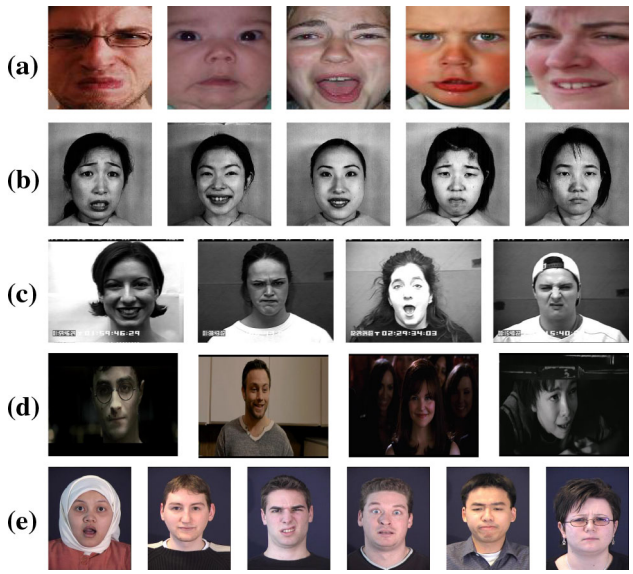


Fig. 8 Example images of multi-label databases. **a** RAF-ML and **b** JAFFE, and single-label databases, **c** CK+, **d** SFEW and **e** MMI

much training samples in JAFFE, we adopted fivefold cross-validation for multi-label classification.

6.5.2 Multi-label to Single-Label

When conducting basic expression recognition tasks on the remaining databases, for the lab-controlled databases CK+ and MMI, we followed the subject-independent experimental principle and conducted fivefold cross validation; for the real-world SFEW 2.0, we followed the rule in EmotiW 2015 (Dhall et al. 2015b): 921 images in the training set are used for training and the learned classifier is then tested on 427 images in the validation set. To avoid parameter sensitivity, support vector machine (SVM) with linear kernel implemented by LibSVM (Chang and Lin 2011) was employed as the classifier. Given a training set $\{(x_i, y_i), i = 1, \ldots, n\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$. With the one-against-one strategy, test sample x_i can be classified by minimizing the objective:

$$\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} max(0, 1 - y_i(w^T x_i + b)), \tag{18}$$

where the tunable parameter C is a regularization term to control the geometric margin and over-fitting.

6.5.3 Parameter Optimization

To optimize the hyper-parameters used in Algorithm 3, we first fixed the parameters k and λ for the bi-manifold loss according to the cross-validation results in Sect. 6.3, i.e., $k = 10$ and $\lambda = 0.01$. Then, we optimized the parameter γ for the MMD loss according to the experimental results on different target databases. Specifically, the value of γ are 0.1, 0.2, 0.15 and 0.25 for JAFFE, CK+, SFEW and MMI, respectively. Moreover, for the multi-label domain to single-label domain scenario, the value of η on the entropy loss is set to 0.05.

6.5.4 Experimental Results

Table 5 reports the detailed experimental results of each comparing feature on JAFFE. And Table 6 reports the performance of our methods on single-label datasets compared with other state-of-the-art results. In Table 6(b), the "SFEW best (Kim et al. 2016)", "SFEW second (Yu and Zhang 2015)" and "SFEW third (Ng et al. 2015)" indicate the best single model result of the 1st, the 2nd and the 3rd participator in EmotiW 2015, respectively. Note that, these participators all trained their model with auxiliary supervised data from SFEW. However, in our deep models, RAF-ML is the only training data with supervised information. For fair comparison, we further trained our network with fine-tuning on the training set of SFEW. The experimental results of DBM-CNN and DBM-DACNN with fine-tuning on SFEW are 52.61% and 54.81%, which outperform the state-of-the-arts method obtained.

From the comparison results in Tables 5 and 6, we can make the following observations. First, DBM-CNN features learned from RAF-ML can achieve comparable performance on other different expression databases. This indicates that the RAF-ML can serve as a 'generic' database containing a great diversity of training data for facial expression analysis. And the learned features can benefit from it due to the particular emotion intensity information it provides. Second, with cross-domain adaption technique that models the biases between different databases, DBM-DACNN can help enhance the performance for this cross-dataset task. Third, by making the learned label distribution present a single peak value, additionally implementing DBM-DACNN with the entropy loss further improves facial expression recognition performance on the single-label target datasets. Under the circumstance, the proposed network can be used as an efficient and effective feature extraction tool for a wide range of facial expression recognition tasks, not only multi-label emotion recognition but also basic emotion recognition, even without the ground truth from target data.

7 Conclusions

In this work, a new real-world facial expression database, RAF-ML, is presented to explore and address challenges of multi-label expression recognition in the wild. Firstly, large number of images have been downloaded from the Internet using emotion-related keywords. Then, a crowd-sourced

Table 5 Experimental results of comparing features on **JAFFE**

Evaluation criterion	Comparing features	Algorithm			
		CLR	RAkEL	MLkNN	LIFT
Hamming loss↓	LBP	0.298	0.236	0.231	0.225
	HOG	0.221	0.183	0.239	0.197
	AlexNet	0.238	0.211	0.212	0.193
	VGG	0.226	0.186	0.236	0.220
	baseDCNN	0.173	0.148	0.202	0.164
	DBM-CNN	0.168	0.146	0.169	0.152
	DBM-DACNN	**0.162**	**0.142**	**0.161**	**0.147**
One-error↓	LBP	0.250	0.298	0.297	0.273
	HOG	0.244	0.234	0.313	0.248
	AlexNet	0.249	0.223	0.223	0.209
	VGG	0.271	0.271	0.323	0.321
	baseDCNN	0.191	0.154	0.175	0.163
	DBM-CNN	0.132	0.117	0.138	0.126
	DBM-DACNN	**0.124**	**0.109**	**0.127**	**0.113**
Coverage↓	LBP	1.749	2.101	1.866	1.756
	HOG	1.696	1.681	1.975	1.802
	AlexNet	1.689	1.854	1.685	1.625
	VGG	1.801	1.718	1.923	1.852
	baseDCNN	1.521	1.436	1.563	1.473
	DBM-CNN	1.345	1.399	1.415	1.401
	DBM-DACNN	**1.236**	**1.247**	**1.369**	**1.341**
Ranking loss↓	LBP	0.149	0.227	0.178	0.176
	HOG	0.142	0.147	0.200	0.164
	AlexNet	0.144	0.176	0.145	0.142
	VGG	0.162	0.153	0.198	0.185
	baseDCNN	0.109	0.096	0.119	0.106
	DBM-CNN	0.078	0.084	0.089	0.075
	DBM-DACNN	**0.062**	**0.075**	**0.082**	**0.072**
Average precision↑	LBP	0.809	0.757	0.783	0.793
	HOG	0.822	0.832	0.766	0.802
	AlexNet	0.822	0.813	0.831	0.833
	VGG	0.795	0.817	0.766	0.802
	baseDCNN	0.863	0.884	0.857	0.872
	DBM-CNN	0.901	0.899	0.891	0.896
	DBM-DACNN	**0.936**	**0.913**	**0.906**	**0.911**

For each evaluation criterion, "↓" indicates the smaller the better while "↑" indicates the bigger the better. The best performance on each evaluation criterion is highlighted in bold face

Table 6 Comparison results of our models and other state-of-the-art methods on three basic expression databases CK+, SFEW 2.0 and MMI

Methods	Accuracy (%)
(a) CK+	
CSPL (Zhong et al. 2012)	88.89
FP+SAE (Lv et al. 2014)	91.11
AUDN (Liu et al. 2013)	92.05
AURF (Liu et al. 2013)	92.22
3DCNN-DAP (Liu et al. 2014a)	92.4
Inception (Mollahosseini et al. 2016)	93.2
Dis-ExpLet (Liu et al. 2016)	95.1
DBM-CNN	94.27
DBM-DACNN without entropy loss	95.18
DBM-DACNN with entropy loss	**96.46**
(b) SFEW 2.0	
DL-GPLVM (Eleftheriadis et al. 2015a)	24.70
AUDN (Liu et al. 2013)	26.14
STM-ExpLet (Liu et al. 2014b)	31.73
Inception (Mollahosseini et al. 2016)	47.7
SFEW third (Ng et al. 2015)	48.5
SFEW second (Yu and Zhang 2015)	52.29
SFEW best (Kim et al. 2016)	52.5
DBM-CNN	50.12 (52.61)
DBM-DACNN without entropy loss	51.46 (53.79)
DBM-DACNN with entropy loss	52.33 (**54.81**)
(c) MMI	
3DCNN-DAP (Liu et al. 2014a)	63.4
DTAGN (Jung et al. 2015)	70.24
CSPL (Zhong et al. 2012)	73.53
AUDN (Liu et al. 2013)	74.76
STM-ExpLet (Liu et al. 2014b)	75.12
F-Bases (Sariyanidi et al. 2017)	75.12
Inception (Mollahosseini et al. 2016)	77.6
Dis-ExpLet (Liu et al. 2016)	77.6
DBM-CNN	78.61
DBM-DACNN without entropy loss	78.90
DBM-DACNN with entropy loss	**79.25**

Bold values indicate the best recognition accuracy of each database

For the last three methods on (b) SFEW 2.0 database, the accuracy inside the parenthesis is achieved by fine-tuning the corresponding methods on the training set of SFEW additionally

label annotation along with an optimization algorithm for crowdsourcing was leveraged to pick out images with multi-label expressions. Focusing on the ambiguity and continuity of blended expressions, a novel deep manifold feature model, DBM-CNN, has been proposed which efficiently considers both feature and label manifold information. Furthermore, domain adaption technique was employed for DBM-CNN to adapt the deep learned features from RAF-ML to other diverse databases. In contrast with most previous studies conducted on controlled data with archetypal emotions, this paper has explored and addressed some of the challenges faced when studying non-basic emotions in the wild. Experimental results on RAF-ML and other related databases demonstrate that our method has the capability of learning more discriminative feature for a wide range of expression analysis tasks. We hope that the release of this database will encourage more researches on multi-label facial expression recognition in the wild .

Acknowledgements The funding was provided by National Natural Science Foundation of China (Grant Nos 61573068, 61471048), Beijing Nova Program (Grant No Z161100004916088).

References

Anitha, C., Venkatesha, M., & Adiga, B. S. (2010). A survey on facial expression databases. *International Journal of Engineering Science and Technology*, *2*(10), 5158–5174.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*, 27:1–27:27. Software Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chang, Y., Hu, C., & Turk, M. (2004). Probabilistic expression analysis on manifolds. In *Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on* (Vol. 2, pp. II–II). IEEE.

Chen, J., Liu, X., Tu, P., & Aragones, A. (2013). Learning person-specific models for facial expression and action unit recognition. *Pattern Recognition Letters*, *34*(15), 1964–1970.

Chu, W. S., De la Torre, F., & Cohn, J. F. (2013). Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3515–3522).

Cour, T., Sapp, B., & Taskar, B. (2011). Learning from partial labels. *Journal of Machine Learning Research*, *12*(May), 1501–1536.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human–computer interaction. *IEEE Signal Processing Magazine*, *18*(1), 32–80.

Csurka, G. (2017). *Domain adaptation for visual applications: A comprehensive survey*. CoRR, arXiv:1702.05374.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In: *CVPR, on* (Vol. 1, pp. 886–893). IEEE.

Dhall, A., Goecke, R., & Gedeon, T. (2015a). Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing*, *6*(1), 13–26.

Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., & Gedeon, T. (2015b). Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 423–426). ACM.

Ding, X., Chu, W. S., De la Torre, F., Cohn, J. F., & Wang, Q. (2013). Facial action unit event detection by cascade of tasks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2400–2407).

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML* (pp. 647–655).

Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, *111*(15), E1454–E1462.

Ekman, P., & Friesen, W. V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues. Journal of Personality* (p. 212). Cambridge, MA: Malor Books.

Ekman, P., Friesen, W. V., & Ellsworth, P. (2013). *Emotion in the human face: Guidelines for research and an integration of findings*. Amsterdam: Elsevier.

Ekman, P., & Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford: Oxford University Press.

Ekman, P., & Scherer, K. (1984). Expression and the nature of emotion. *Approaches to Emotion*, *3*, 19–344.

Eleftheriadis, S., Rudovic, O., & Pantic, M. (2015a). Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE Transactions on Image Processing*, *24*(1), 189–204.

Eleftheriadis, S., Rudovic, O., & Pantic, M. (2015b). Multi-conditional latent variable model for joint facial action unit detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 3792–3800).

Fabian Benitez-Quiroz, C., Srinivasan, R., & Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5562–5570).

Fürnkranz, J., Hüllermeier, E., Mencía, E. L., & Brinker, K. (2008). Multilabel classification via calibrated label ranking. *Machine Learning*, *73*(2), 133–153.

Gao, B. B., Xing, C., Xie, C. W., Wu, J., & Geng, X. (2017). Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, *26*(6), 2825–2838.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012a). A kernel two-sample test. *Journal of Machine Learning Research*, *13*(Mar), 723–773.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., & Sriperumbudur, B. K. (2012b). Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems* (pp. 1205–1213).

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on* (Vol. 2, pp. 1735–1742). IEEE.

Hassin, R. R., Aviezer, H., & Bentin, S. (2013). Inherently ambiguous: Facial expressions of emotions, in context. *Emotion Review*, *5*(1), 60–65.

He, X., & Niyogi, P. (2004). Locality preserving projections. In *Advances in neural information processing systems* (pp. 153–160).

Hou, P., Geng, X., & Zhang, M. L. (2016). Multi-label manifold learning. In *AAAI* (pp. 1680–1686).

Huang, S. J., Zhou, Z. H., & Zhou, Z. (2012). Multi-label learning by exploiting label correlations locally. In *AAAI* (pp. 949–955).

Inc. M. (2013). Face++ research toolkit. www.faceplusplus.com.

Izard, C. E. (1972). Anxiety: A variable combination of interacting fundamental emotions. In *Anxiety: Current trends in theory and research* (Vol. 1, pp. 55–106).

Izard, C. E. (2013). *Human emotions*. Berlin: Springer.

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*(19), 7241–7244.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on multimedia* (pp. 675–678). ACM.

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).

Kim, B. K., Roh, J., Dong, S. Y., & Lee, S. Y. (2016). Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User*, *Interfaces*, 1–17.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Li, S., & Deng, W. (2018). *Deep facial expression recognition: A survey*. CoRR, arXiv:1804.08348.

Li, S., & Deng, W. (2019). 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., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2584–2593). IEEE.

Liu, M., Li, S., Shan, S., & Chen, X. (2013). Au-aware deep networks for facial expression recognition. In *Automatic face and gesture recognition (FG), 2013 10th IEEE international conference and workshops on* (pp. 1–6). IEEE.

Liu, M., Li, S., Shan, S., Wang, R., & Chen, X. (2014a). Deeply learning deformable facial action parts model for dynamic expression analysis. In *Asian conference on computer vision* (pp. 143–157). Berlin: Springer.

Liu, M., Shan, S., Wang, R., & Chen, X. (2014b). Learning expression lets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1749–1756).

Liu, M., Shan, S., Wang, R., & Chen, X. (2016). Learning expressionlets via universal manifold model for dynamic facial expression recognition. *IEEE Transactions on Image Processing*, *25*(12), 5920–5932.

Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *International conference on machine learning* (pp. 97–105).

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 *CVPRW, on* (pp. 94–101). IEEE.

Lv, Y., Feng, Z., & Xu, C. (2014). Facial expression recognition via deep learning. In *Smart computing (SMARTCOMP), 2014 international conference on* (pp. 303–308). IEEE.

Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding facial expressions with Gabor wavelets. In *Automatic face and gesture recognition, 1998. Proceedings. Third IEEE international conference on* (pp. 200–205). IEEE.

Miao, Y. Q., Araujo, R., & Kamel, M. S. (2012) Cross-domain facial expression recognition using supervised kernel mean matching. In *Machine learning and applications (ICMLA), 2012 11th international conference on, IEEE* (Vol. 2, pp. 326–332).

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

Ng, H. W., Nguyen, V. D., Vonikakis, V., & Winkler, S. (2015). Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 443–449). ACM.

Nummenmaa, T. (1988). The recognition of pure and blended facial expressions of emotion from still photographs. *Scandinavian Journal of Psychology*, *29*(1), 33–47.

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987.

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.

Patel, V. M., Gopalan, R., Li, R., & Chellappa, R. (2015). Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, *32*(3), 53–69.

Plutchik, R. (1991). *The emotions*. Lanham: University Press of America.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, *76*(5), 805.

Sariyanidi, E., Gunes, H., & Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(6), 1113–1133.

Sariyanidi, E., Gunes, H., & Cavallaro, A. (2017). Learning bases of activity for facial expression recognition. *IEEE Transactions on Image Processing*, *26*(4), 1965–1978. https://doi.org/10.1109/TIP.2017.2662237.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. (pp. 815–823).

Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 806–813).

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. CoRR, arXiv:1409.1556.

Tomkins, S. S. (1963). *Affect imagery consciousness: Volume II: The negative affects* (Vol. 2). Berlin: Springer.

Tsoumakas, G., & Katakis, I. (2006). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, *3*(3), 1–13.

Tsoumakas, G., & Vlahavas, I. (2007). Random k-labelsets: An ensemble method for multilabel classification. In *Machine learning: ECML 2007* (pp. 406–417). Berlin: Springer.

Valstar, M., & Pantic, M. (2010). Induced disgust, happiness and surprise: An addition to the MMI facial expression database. In *Proceedings of the 3rd international workshop on EMOTION (satellite of LREC): Corpora for research on emotion and affect* (p. 65).

Viola, P., & Jones, M. (2001) Rapid object detection using a boosted cascade of simple features. In *CVPR, on* (Vol. 1, pp. 1–511). IEEE.

Wang, S., Liu, Z., Wang, J., Wang, Z., Li, Y., Chen, X., et al. (2014). Exploiting multi-expression dependences for implicit multi-emotion video tagging. *Image and Vision Computing*, *32*(10), 682–691.

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*. Berlin: Springer (pp. 499–515).

Whitehill, J., Wu, T. F., Bergsma, J., Movellan, J. R., & Ruvolo, P. L. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems* (pp. 2035–2043).

Xing, C., Geng, X., & Xue, H. (2016). Logistic boosting regression for label distribution learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4489–4497).

Xiong, X., & De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 532–539).

Yan, H., Ang, M. H., & Poo, A. N. (2011). Cross-dataset facial expression recognition. In *Robotics and automation (ICRA), 2011 IEEE international conference on* (pp. 5985–5990). IEEE.

Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. J. (2006). A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on* (pp. 211–216). IEEE.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320–3328).

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

Zen, G., Porzi, L., Sangineto, E., Ricci, E., & Sebe, N. (2016). Learning personalized models for facial expression analysis and gesture recognition. *IEEE Transactions on Multimedia*, *18*(4), 775–788.

Zeng, J., Chu, W. S., De la Torre, F., Cohn, J. F., & Xiong, Z. (2015). Confidence preserving machine for facial action unit detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 3622–3630).

Zhang, M. L., & Wu, L. (2015). Lift: Multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(1), 107–120.

Zhang, M. L., & Yu, F. (2015). Solvingthe partial label learning problem: An instance-based approach. In *IJCAI* (pp. 4048–4054).

Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038–2048.

Zhang, M. L., & Zhou, Z. H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *26*(8), 1819–1837.

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

Zhao, K., Zhang, H., Ma, Z., Song, Y. Z., & Guo, J. (2015). Multi-label learning with prior knowledge for facial expression analysis. *Neurocomputing*, *157*, 280–289.

Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., & Metaxas, D. N. (2012). Learning active facial patches for expression analysis. In *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on* (pp. 2562–2569). IEEE.

Zhou, Y., Xue, H., & Geng, X. (2015). Emotion distribution recognition from facial expressions. In *Proceedings of the 23rd ACM international conference on multimedia* (pp. 1247–1250). ACM.

Zhu, R., Sang, G., & Zhao, Q. (2016). Discriminative feature adaptation for cross-domain facial expression recognition. In *Biometrics (ICB), 2016 international conference on* (pp. 1–7). IEEE.

Zong, Y., Huang, X., Zheng, W., Cui, Z., & Zhao, G. (2017). Learning a target sample re-generator for cross-database micro-expression recognition. In *Proceedings of the 2017 ACM on multimedia conference* (pp. 872–880). ACM.