Computational Emotion Analysis From Images: Recent Advances and Future Directions

Sicheng Zhao, Quanwei Huang, Youbao Tang, Xingxu Yao, Jufeng Yang, Guiguang Ding, and Björn W. Schuller

1 Introduction

With the rapid development and popularity of social networks, such as Twitter[1] and Sina Weibo.[2] people tend to express and share their opinions and emotions online using text, images, and videos. Understanding the information contained in the increasing repository of data is of vital importance to behavior sciences (Pang & Lee 2008), which aim to predict human decision making and enable wide applications, such as mental health evaluation (Guntuku et al. 2019), business recommendation (Pan et al. 2014), opinion mining (Tumasjan et al. 2010), and entertainment assistance (Zhao et al. 2020).

Analyzing media data on an affective (emotional) level belongs to affective computing, which is defined as "*the computing that relates to, arises from, or influences emotions*" (Picard 2000). The importance of emotions has been emphasized for decades since Minsky introduced the relationship between intelligence and emotion (Minsky 1986). One famous claim is "*The question is not whether*

[1] https://twitter.com.

[2] http://www.weibo.com.

S. Zhao (✉) · Q. Huang · G. Ding
Tsinghua University, Beijing, China

Y. Tang
PAII Inc., Palo Alto, CA, USA

X. Yao · J. Yang
Nankai University, Tianjin, China

B. W. Schuller
GLAM, Imperial College London, London, UK

intelligent machines can have any emotions, but whether machines can be intelligent without emotions." Based on the types of media data, the research on affective computing can be classified into different categories, such as text (Giachanou & Crestani 2016; Zhang et al. 2018), image (Zhao et al. 2018), speech (Schuller 2018), music (Yang & Chen 2012), facial expression (Li & Deng 2020), video (Wang & Ji 2015; Zhao et al. 2020), physiological signals (Alarcao & Fonseca 2019), and multi-modal data (Soleymani et al. 2017; Poria et al. 2017; Zhao et al. 2019).

The adage "*a picture is worth a thousand words*" indicates that images can convey rich semantics. Therefore, images are used as an important channel to express emotions. Image emotion analysis (IEA) has recently been paid much attention. As compared to analyzing the images' cognitive aspect that is related with objective content (Hanjalic 2006), such as object classification and semantic segmentation, IEA focuses on understanding what emotions can be induced by the images in viewers. The challenges of affective gap and perception subjectivity (Zhao et al. 2018) make IEA a difficult task.

In this chapter, we concentrate on introducing recent advances on IEA—especially our recent efforts from a computational perspective and on suggesting future research directions. First, we briefly introduce some popular emotion representation models from psychology in Sect. 2, define corresponding key computational problems, and provide some representative supervised frameworks in Sect. 3. Second, we introduce the major challenges in IEA in Sect. 4. Third, we present some representative methods on different computational components, such as emotion feature extraction in Sect. 5 and supervised classifier learning as well as domain adaptation in Sect. 6. Then, we introduce some typical datasets for IEA evaluation in Sect. 7 and investigate the performances of different features and classifiers on these datasets in Sect. 8, as emotions can be conveyed by various features, as shown in Fig. 1. Finally, we give a discussion on what questions are still open and provide some suggestions for future research in Sect. 9.

2 Emotion Representation Models from Psychology

Psychologists have proposed different theories to explain the what, how, and why behind human emotions (Plutchik & Kellerman 2013). For example, the James-Lange theory suggests that emotions occur as a result of physiological reactions to events; the Cognitive Appraisal theory claims that the sequence of events first involves a stimulus, followed by thought, which then leads to the simultaneous physiological response and emotion. Some other emotion theories include the Evolutionary theory, the Cannon-Bard theory, the Schachter-Singer Theory, and the Facial-Feedback theory (Plutchik & Kellerman 2013).

Besides emotion, several other concepts (e. g., affect, sentiment, feeling, and mood) are also widely used in psychology. The difference or correlation of these concepts can be found in Munezero et al. (2014). In this chapter, we focus on a computational perspective and do not distinguish them clearly, except sentiment

Fig. 1 The emotions conveyed by different kinds of images are correlated with different features Zhao et al. (2014): (**a**) Aesthetic features (low saturation, cool color, low color difference); (**b**) Attributes (snow, skiing); (**c**) Semantic concepts described by adjective noun pairs (broken car); (**d**) Facial expressions (happiness). (**a**) Fear. (**b**) Excitement. (**c**) Sadness. (**d**) Contentment

for positive/negative/neutral categories and emotion for more fine-grained definitions. Another relevant concept is about expected, induced, or perceived emotion. Expected emotion is the emotion that the image creator intends to make people feel, perceived emotion is what people perceive as being expressed, while induced/felt emotion is the actual emotion that is felt by a viewer. Interested readers can refer to Juslin and Laukka (2004) for more details. Unless otherwise specified, the emotion focused in this chapter is about induced emotion because of the dataset construction process.

To quantitatively measure emotion, psychologists have mainly employed two types of emotion representation models, categorical emotion states (CES) and dimensional emotion space (DES) (Zhao et al. 2018). For CES, a set of preselected categories is used to define emotions. Some popular CES models include binary sentiment (positive and negative, sometimes including neutral), Ekman's six basic emotions (happiness, surprise and *negative* anger, disgust, fear, and sadness) (Ekman 1992), and Mikels's eight emotions (amusement, anger, awe, contentment, disgust, excitement, fear, and sadness) (Mikels et al. 2005). More diverse and fine-grained emotion categories are being increasingly considered. In Plutchik's emotion model (Plutchik 1980), each basic emotion category (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) is organized into three intensities. For example, the three intensities from low to high for surprise are distraction$\longrightarrow$surprise$\longrightarrow$amazement. Parrott represents emotions with a three-level hierarchy, i. e., primary (positive and negative), secondary (anger, fear, joy, love, sadness, and surprise), and tertiary (25 fine-grained categories) (Parrott 2001). For DES, a 2D, 3D, or higher dimensional Cartesian space is employed to

represent emotions, such as valence-arousal-dominance (VAD) (Schlosberg 1954) and activity-temperature-weight (Lee & Park 2011). VAD is the most widely used DES model, where 'V' represents the pleasantness ranging from positive to negative, 'A' represents the intensity of emotion ranging from excited to calm, and 'D' represents the degree of control ranging from controlled to in control.

Intuitively, CES models are easy for users to understand, but limited emotion categories cannot well reflect the complexity and subtlety of emotions. Further, psychologists have not reached a consensus on how many categories should be included. Theoretically, all emotions can be measured as different coordinate points in the continuous Cartesian space. However, such absolute continuous values are difficult for non-experts to understand. Specifically, CES can be transformed to DES but not all Cartesian points can correspond to detailed categories (Alarcão & Fonseca 2018). For example, fear is often related to negative valence, high arousal, and low dominance. In this chapter, the employed CES models mainly include binary sentiment and Mikels's eight emotions, and VAD is employed as the DES model.

3 Key Computational Problems and Supervised Frameworks

Based on different emotion representation models, we can perform different IEA tasks: classification/retrieval based on CES, and regression/retrieval based on DES. Current methods mainly employ supervised methods with the help of available labeled datasets. In this section, we will define the key computational problems and provide representative supervised frameworks.

3.1 Emotion Classification and Regression

Suppose all images in the dataset are grouped into K emotion categories, then emotion prediction can be conceived as a multi-class classification problem. Based on the model trained on given training samples, an emotion category that is most likely evoked in humans is assigned to a test image. Suppose we have N training images $\{(\mathbf{x}_i, y_i)_{i=1}^N\}$, where $y_i \in \{1, 2, \cdots, K\}$. Let $g_\mu(\mathbf{x})$ denote the feature extractor of image $\mathbf{x}$, and then our goal is to learn some model $h_\theta(g_\mu(\mathbf{x})) : g_\mu(\mathbf{x}) \rightarrow y$ that maps image features $g_\mu(\mathbf{x})$ to emotion labels y, where μ and θ are parameters. Usually, the learning process is transformed to a parameter optimization problem, which can be defined as

$$\begin{aligned} J(\omega, \theta, \mu) &= \sum_{i=1}^{N} f_\omega(h_\theta(g_\mu(\mathbf{x}_i), y_i), \\ [\omega^*, \theta^*, \mu^*] &= \arg\min_{\omega,\theta,\mu} J(\omega, \theta, \mu), \end{aligned} \tag{1}$$

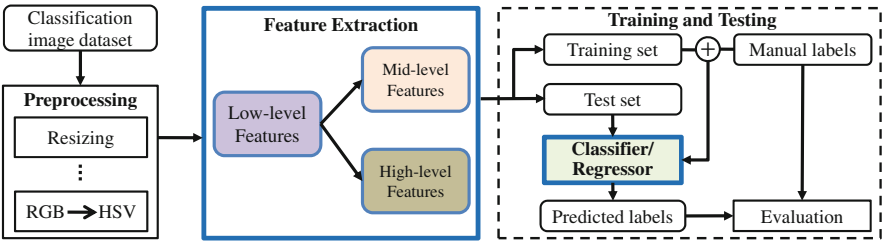

Fig. 2 Commonly used supervised framework of affective image classification and regression. The key components researchers have been studying lie in the solid blue rectangles

where $f_\omega(.,.)$ is a function with parameters ω to compute the loss function $J(\omega, \theta, \mu)$ between the predicted labels and the ground truth, and arg min is the argument of the minimum. Once we work out μ and θ, given a test image $\mathbf{x}_{te}$, we can obtain the prediction label $h_\theta(g_\mu(\mathbf{x}_{te}))$.

Emotion regression assumes that emotions are represented by continuous dimensional values instead of discrete emotion labels, i. e., y is continuous. Except this, the learning process of emotion regression is analogous to emotion classification.

The commonly used supervised framework of affective image classification and regression is shown in Fig. 2. Firstly, some preprocessing is done to 'normalize' the images. Then, different features are extracted for each image, which presents the core of image emotion analysis and will be described in detail. The dataset is split into a training set and a test set. A classifier or regressor is trained using the training set along with the emotion labels based on certain learning models. The images in the test set are then automatically classified by the trained classifier or regressed by the trained regressor. The assigned emotion labels are compared with the ground truth to evaluate the classification or regression performance.

3.2 Emotion Retrieval

Affective image retrieval, firstly named emotional semantic image retrieval (Wang & He 2008), involves searching for images that express similar emotions to the query image. Affective image retrieval can be formalized as a reranking problem to ensure that the top ranked images are the ones emotionally similar to the query image.

Suppose the features and emotion label of a given query image $\mathbf{x}_q$ are $g_\mu(\mathbf{x}_q)$ and y_q, and in the dataset there are N_s emotionally similar images, in which the features and labels of the ith image are $\mathbf{x}_i^s$ and y_i^s, where $y_i^s == y_q, i = 1, 2, \cdots, N_s$, and N_d emotionally different images, in which the features and labels of the jth image are $\mathbf{x}_j^d$ and y_j^d, where $y_j^d \neq y_q, j = 1, 2, \cdots, N_d$. Then, our goal is to minimize

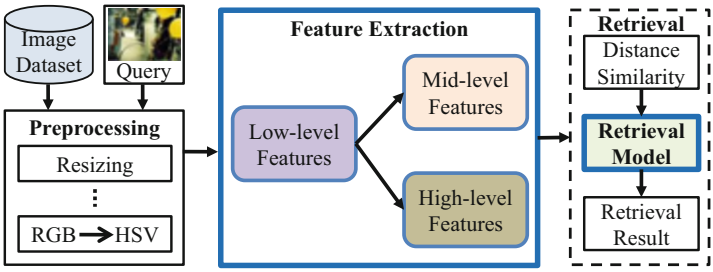

Fig. 3 Commonly used supervised framework of affective image retrieval. The key components researchers have been studying lie in the solid blue rectangles

the distance between the query image and the N_s positive images and maximize the distance between the query image and the N_d negative images:

$$\begin{aligned} J_s(\theta, \mu) &= \sum_{i=1}^{N_s} h_\theta(D(g_\mu(\mathbf{x}_i^s), g_\mu(\mathbf{x}_q)), \\ J_d(\theta, \mu) &= \sum_{j=1}^{N_d} h_\theta(D(g_\mu(\mathbf{x}_j^d), g_\mu(\mathbf{x}_q)), \\ J(\omega, \theta, \mu) &= f_\omega(J_s(\theta, \mu), J_d(\theta, \mu)), \\ [\omega^*, \theta^*, \mu^*] &= \arg\min_{\omega,\theta,\mu} J(\omega, \theta, \mu), \end{aligned} \tag{2}$$

where $D(.,.)$ is a distance function to compute the distance between two feature vectors, such as the Minkowski-form distance and the Mahalanobis distance, $h_\theta(.)$ is a function with parameters θ to compute a cost of the query image and the image in the dataset, $f_\omega(.,.)$ is a function with parameters ω to compute the total cost $J(\omega, \theta, \mu)$ between the positive cost $J_s(\theta, \mu)$ and the negative cost $J_d(\theta, \mu)$. Once we work out μ and θ, we can get the retrieval results by sorting the cost.

The commonly used supervised framework of affective image retrieval is shown in Fig. 3. The preprocessing and feature extraction parts are similar to the related parts in emotion classification and regression. The distance or similarity is computed between the features of the query image and each image in the dataset. Through some retrieval model, we sort the distance or similarity and obtain the retrieval results, which are compared with the ground truth for evaluation.

4 Major Challenges

Affective Gap The affective gap is one main challenge for IEA, which is defined as the inconsistency between extracted low-level features and induced emotions (Hanjalic 2006; Zhao et al. 2018). As compared to the semantic gap in computer vision,

i. e., the discrepancy between the limited descriptive power of low-level visual features and the richness of user semantics (Smeulders et al. 2020; Liu et al. 2007), the affective gap is even more challenging. Bridging the semantic gap cannot guarantee bridging the affective gap. For example, images containing a barking dog and a loving dog are both about dogs but obviously induce different emotions. To bridge the affective gap, the main efforts have been focusing on designing and extracting discriminative emotion features, ranging from the early hand-crafted features to more recent deep ones. Based on these features, a dominant emotion category (DEC) is assigned to an image by traditional single-label learning-based methods.

Perception Subjectivity Emotion is a highly subjective and complex variable. Different viewers may perceive totally different emotions to the same image, which is influenced by many factors, such as culture, education, personality, and environment (Zhao et al. 2018). For example, for a sudden heavy snow, some may feel excitement to see such rare natural scenes, some may feel sadness because the planned activities have to be cancelled, some may feel amusement since they can build a snowman, etc. For the subjectivity challenge, one direct and intuitive solution is to predict emotions for each viewer via personalized learning models (Zhao et al. 2018). When a large number of viewers are involved, we can assign the image with multiple emotion labels via multi-label learning methods. Since the importance or extent of different labels is actually unequal, predicting the probability distribution of emotions, either discrete (Yang et al. 2017; Zhao et al. 2020) or continuous (Zhao et al. 2017), would make more sense.

Label Noise and Absence Recent deep learning based IEA methods have achieved state-of-the-art performances with the help of large-scale labeled training data. However, in real applications, it is expensive and time-consuming and even impossible to obtain sufficient data with emotion labels to train a deep model. It would be more practical if we can deal with the situation that there are only few or even no emotion labels. We can conduct unsupervised/weakly supervised learning (Wei et al. 2020) and few/zero shot learning (Zhan et al. 2019). One might consider leveraging the large amount of weakly-labeled web images (Wei et al. 2020). Since the associated tags might contain noise that is unrelated to emotion and even to visual semantics, filtering such automatic labels is necessary. Another possible solution is to transfer the well-learned model on one labeled source domain to another unlabeled or sparsely labeled target domain. Direct transfer often results in obvious performance decay, because of the influence of domain shift (Zhao et al. 2021), i. e., the joint distribution of images and emotion labels are different across domains. To bridge the domain shift challenge, we can employ domain adaptation and domain generalization techniques (Zhao et al. 2021).

5 Emotion Features

In this section, we summarize the features that have been widely extracted for IEA, including both hand-crafted and deep features. We first give an brief overview and then introduce some representative ones especially our recent work.

5.1 Hand-Crafted Features

Overview Early efforts on IEA mainly focused on hand-crafting features from different levels. *Low-level features* are used in the earliest IEA methods, which suffer from large affective gap and low interpretability. Some generic features from computer vision, such as Gabor, HOG, and GIST, are directly used in the IEA task (Yanulevskaya et al. 2008). Some specific features derived from elements of art, including color and texture, are implemented (Machajdik & Hanbury 2010). Low-level color features include mean saturation and brightness, vector based mean hue, emotional coordinates (pleasure, arousal and dominance) based on brightness and saturation, colorfulness and color names. Low-level texture features include Tamura texture, Wavelet textures, and gray-level co-occurrence matrix (GLCM) based texture (Machajdik & Hanbury 2010). Low-level shape features, including line segments, angles, continuous lines, and curves, are designed in Lu et al. (2012). As compared to low-level features, *mid-level features* are more interpretable, semantic, and relevant to emotions. Different types of attributes people use to describe scenes, such as materials, surface properties, functions or affordances, spatial envelope attributes, and object presence are modeled (Yuan et al. 2013). Features inspired from principles of art, such as symmetry, emphasis, harmony, and variety, are specially designed (Zhao et al. 2014). *High-level features* describe the detailed content in an image through which viewers can easily understand the semantics and evoked emotions. Some representative high-level features include adjective noun pairs detected by SentiBank Borth et al. (2013) and recognized facial expressions (Yang et al. 2010).

Mid-level Principles-of-art Based Emotion Features The principles of art are defined as the rules, tools, or guidelines of arranging and orchestrating the elements of art in an artwork. They consider various artistic aspects including balance, emphasis, harmony, variety, gradation, movement, rhythm, and proportion (Zhao et al. 2014). The comparison of elements of art and principles of art is shown in Fig. 4. Six principles of art are formulated and implemented systematically in Zhao et al. (2014) based on related art theory and multimedia research. Totally, a 165 dimensional feature can be obtained for each image. For example, emphasis, also known as contrast, is used to stress the difference of certain elements, which can be accomplished by using sudden and abrupt changes in elements. Itten color contrast, which is defined to coordinate colors using the hue's contrasting properties, is implemented (Zhao et al. 2014), including contrast of saturation, contrast of light

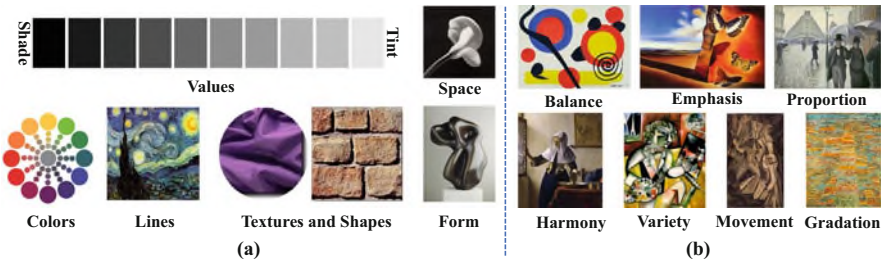

Fig. 4 Illustration of artistic elements and artistic principles, which are designed as low-level and mid-level emotion features. (**a**) Elements of art. (**b**) Principles of art

and dark, contrast of extension, contrast of complements, contrast of hue, contrast of warm and cold, and simultaneous contrast. The results show that principles of art features are more correlated with emotions than elements of art (Zhao et al. 2014). For example, images with high balance and harmony values tend to express positive emotions.

High-Level Adjective Noun Pairs The adjective noun pairs (ANPs) are detected by a large detector library SentiBank (Borth et al. 2013), which is trained using GIST, a 3×256 dimension color histogram, a 53 dimensional LBP descriptor, a Bag-of-Words quantized descriptor using a 1000 word dictionary with a 2-layer spatial pyramid and max pooling, and a 2000 dimensional attribute on about 500k images downloaded from Flickr. Liblinear support vector machine (SVM) (Fan et al. 2008) is used as classifier and early fusion is adopted. The advantages of ANP are that it turns a neutral noun into an ANP with strong emotions and makes the concepts more detectable, as compared to nouns and adjectives, respectively. Finally, a 1200 dimensional double vector representing the probability of the ANPs is obtained.

5.2 Deep Features

Overview With the development of deep learning, especially convolutional neural networks (CNNs), learning-based deep features have been widely employed with superior performances as compared to hand-crafted ones. *Global features* are directly extracted from the whole images. One direct and intuitive method is to employ the output of the last few fully connected (FC) layers as deep features, using either pretrained or finetuned CNN models (Xu et al. 2014; Chen et al. 2015; You et al. 2016). The last few FC layers correspond to high-level semantic features, which might be not enough to represent emotions, especially for abstract images. Therefore, some methods try to extract multi-level deep features (Rao et al. 2020; Zhu et al. 2017; Yang et al. 2018). For example, three parallel networks, namely an Alexnet, an aesthetics CNN, and a texture CNN, are trained with different levels of

image patches as input. Deep representations at three levels, i. e., image semantics, image aesthetics, and low-level visual features are extracted. The features from different layers in CNNs are extracted as multi-level representations, which are fed into a bidirectional gated recurrent unit model to exploit the dependency among different levels of features (Zhu et al. 2017). The above methods treat different regions of an image equally. Based on the fact that some regions can determine the emotion of an image while the other regions do not help much and might even reverse, some recent methods focus on extracting *local features* that are more discriminative for IEA (You et al. 2017; She et al. 2020; Zhao et al. 2019; Yao et al. 2020).

Weakly Supervised Coupled Networks (WSCNet) WSCNet contains two branches for joint emotion detection and classification (She et al. 2020). One is the detection branch which is designed to generate region proposals that evoke emotion. A soft sentiment map is generated by a cross-spatial pooling strategy to summarize all the information contained in the feature maps for each category. The regions of interest that are informative for classification are highlighted in the sentiment map. The advantage of such setting is that the network can be trained with image-level emotion labels, without requiring time-consuming region-level annotation. The other is the classification branch designed for the emotion classification task by considering both global and local representations. The global features are extracted from a fully convolutional network (FCN), while the local features are obtained by coupling the generated sentiment map in the detection branch with the global features.

Polarity-Consistent Deep Attention Network (PDANet) The feature maps of PDANet from a FCN are fed into two branches (Zhao et al. 2019), as shown in Fig. 5. Each branch is a multi-layer neural network. One is used to estimate the spatial attention to emphasize the emotional semantic-related regions by two 1×1 convolutional layers and a hyperbolic tangent function. The other is used to estimate the channel-wise attention to consider the interdependency between different channels by one 1×1 convolutional layer and a sigmoid function. The attended semantic vectors that capture the global and local information respectively are concatenated as the final feature representations for IEA tasks.

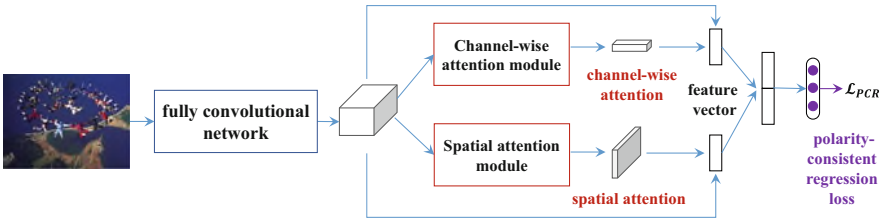

Fig. 5 Overview of the polarity-consistent deep attention network (PDANet) Zhao et al. (2019) to extract attended features for IEA

Attention-Aware Polarity-Sensitive Embedding (APSE) APSE utilizes a hierarchical attention mechanism to learn both polarity and emotion-specific attended representations (Yao et al. 2020). Based on the fact that concrete emotion categories depend on high-level semantic information and that polarity is relevant to low-level features (e. g., color and texture), polarity-specific attention is modeled in lower layers and emotion-specific attention is modeled in higher layers. These two types of attended features are integrated by cross-level bilinear pooling to facilitate the interaction between the information of different levels. After dimensionality reduction and ℓ_2-Normalization, we can obtain the final feature representations.

6 Learning Methods for IEA

In this section, we first summarize the supervised learning methods that have been widely used for emotion classification, regression and retrieval. Then, we introduce some domain adaptation methods.

6.1 Emotion Classification

Shallow Pipeline Based on the modeling process, supervised learning can be classified into generative learning and discriminative learning. Discriminative learning models the conditional distribution of labels y given features $g_\mu(\mathbf{x})$ directly or learns the mappings directly from features $g_\mu(\mathbf{x})$ to labels y. For instance, logistic regression, a binary classification method, models the conditional distribution $p(y|g_\mu(\mathbf{x}); \theta)$ as:

$$h_\theta(g_\mu(\mathbf{x})) = \text{sig}(\theta^T g_\mu(\mathbf{x})), \tag{3}$$

where sig is the sigmoid function $\text{sig}(z) = \dfrac{1}{1+e^{-z}}$ and θ is the vector of parameters. A generalization of logistic regression to multi-class classification is softmax regression. The perceptron learning algorithm 'forces' the output values of logistic regression to be exactly 0 or 1, based on the threshold function:

$$\text{sig}(z) = \begin{cases} 1, & \text{if} z \geq 0, \\ 0, & \text{if} z < 0. \end{cases} \tag{4}$$

Support vector machines (SVM) try to find a decision boundary that maximizes the geometric margin and can be extended with various non-linear kernels.

Generative learning algorithms try to model class priors $p(y)$ and likelihood $p(g_\mu(\mathbf{x})|y)$, and then, the posterior distribution on $p(y|g_\mu(\mathbf{x}))$ can be derived by Bayes rule:

$$p(y|g_\mu(\mathbf{x})) = \frac{p(g_\mu(\mathbf{x})|y)p(y)}{p(g_\mu(\mathbf{x}))}, \tag{5}$$

where $p(g_\mu(\mathbf{x}))$ can be seen as a normalization factor. Gaussian discriminant analysis assumes that $p(g_\mu(\mathbf{x})|y)$ is distributed according to a multivariate Gaussian distribution, which deals with continuous real-valued features. Naive Bayes, which handles discrete values of $g_\mu(\mathbf{x})$, is based on the assumption that the discrete values are conditionally independent given y. When dealing with multi-class classification, it is often formulated as some extensions of binary classification. The prominent formulations include 'one-versus-all' and 'one-versus-one' classification.

Deep Architecture Recent deep learning based emotion classification methods usually employ several fully-connected (FC) layers to minimize the following cross-entropy loss (She et al. 2020):

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{1}_{[k=y_i]}\log p_{i,k}, \tag{6}$$

where K is the number of emotion classes, $\mathbb{1}_{[k=y_i]}$ is a binary indicator, and $p_{i,k}$ is the predicted probability that image i belongs to class k. Directly optimizing the cross-entropy loss might lead some images to be incorrectly classified into categories with opposite polarity. For example, for an image with the emotion "amusement", one model might classify the emotion incorrectly as "sadness" which has an opposite polarity (negative vs. positive). But it is more acceptable if the emotion is classified as "excitement" which has the same polarity (positive). Based on this motivation, a novel polarity-consistent cross-entropy (PCCE) loss is proposed to consider the polarity-emotion hierarchy by increasing the penalty of the predictions that have opposite polarity to the ground truth (Zhao et al. 2020). The PCCE loss is defined as:

$$\mathcal{L}_{PCCE} = -\frac{1}{N}\sum_{i=1}^{N}(1+\lambda(G(\hat{y}_i, y_i)))\sum_{k=1}^{K}\mathbb{1}_{[k=y_i]}\log p_{i,k}, \tag{7}$$

where λ is a penalty coefficient. Similar to the indicator function, $G(.)$ represents whether to add the penalty or not and is defined as:

$$G(\hat{y}, y) = \begin{cases} 1, & \text{if polarity } (\hat{y}) \neq \text{polarity}(y), \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

where polarity(.) is a function that maps an emotion category to its polarity (positive or negative).

6.2 Emotion Regression

In the early shallow pipeline, some commonly used regression methods, including linear regression, support vector regression (SVR), and manifold kernel regression, are employed to predict the average dimensional values. For example, SVR is used in (Lu et al. 2012) to predict emotion scores in the VA space.

Similar to emotion classification, deep learning based emotion regression methods also employ several fully-connected (FC) layers to minimize the following mean squared error (MSE):

$$\mathcal{L}_{reg} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_E}(\hat{y}_i^j - y_i^j)^2, \tag{9}$$

where N_E is the dimension number of the adopted emotion model ($N_E = 3$ for VAD), and y_i^j indicates the emotion label of the j-th dimension for image $\mathbf{x}_i$. Similar to PCCE loss, polarity-consistent regression (PCR) loss is proposed based on the assumption that VAD dimensions can be classified into different polarities (Zhao et al. 2019). The PCR loss is defined as:

$$\mathcal{L}_{PCR} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_E}(\hat{y}_i^j - y_i^j)^2(1 + \lambda G(\hat{y}_i^j, y_i^j)). \tag{10}$$

6.3 Emotion Retrieval

We introduced our work on multi-graph learning (MGL) (Zhao et al. 2014) and attention-aware polarity-sensitive embedding (APSE) (Yao et al. 2020) as shallow and deep methods for emotion retrieval. As a (semi-)supervised learning, MGL is widely used for reranking in various domains. For each feature, we can construct a single graph, where the vertices represent image samples and the edges reflect the similarities between sample pairs. By combining the multiple graphs together in a regularization framework, we can learn the optimized weights of each graph to efficiently explore the complementarity of different features (Zhao et al. 2014).

Besides the polarity and emotion-specific attended representations, APSE also consists of a polarity-sensitive emotion-pair (EP) loss to further exploit the polarity-emotion hierarchy (Yao et al. 2020). Suppose K pairs of convolution features constructed from K different categories are formulated as $\left\{(g_1, g_1^+), \cdots, (g_K, g_K^+)\right\}$,

where g_k and g_k^+ represent the feature representations of anchor point $\mathbf{x}_k$ and positive example $\mathbf{x}_k^+$, respectively, both from the k^{th} category. The EP loss is the combination of inter-polarity loss and intra-polarity loss. Specifically, inter-polarity loss is formulated as:

$$\mathcal{L}_{inter} = \frac{1}{K}\sum_{k=1}^{K}\log(1+\exp(\frac{1}{N_{\mathcal{Q}_k}}\sum_{j\in\mathcal{Q}_k} g_k^\top g_j^+ - \frac{1}{N_{\mathcal{P}_k}}\sum_{j\in\mathcal{P}_k, j\neq k} g_k^\top g_j^+)), \tag{11}$$

where $\mathcal{P}_k$ and $\mathcal{Q}_k$ represent the sets of emotion categories in the same and opposite polarities to the anchor of the k^{th} category, respectively. $N_{\mathcal{P}_k}$ and $N_{\mathcal{Q}_k}$ are the numbers of corresponding categories. The intra-polarity loss that can differentiate similar categories within the same polarity is defined as:

$$\mathcal{L}_{intra} = \frac{1}{K}\sum_{k=1}^{K}\log(1+\sum_{j\in\mathcal{P}_k, j\neq k}\exp(g_k^\top g_j^+ - g_k^\top g_k^+)). \tag{12}$$

6.4 Emotion Distribution Learning

Emotion distribution learning is essentially a regression problem. We can directly employ regression methods to predict the probabilities of each emotion category, but the relationship between different emotion categories is ignored. Shared sparse learning (SSL) is employed to learn the probabilities of different emotion categories simultaneously as a distribution (Zhao et al. 2020). SSL is performed based on two assumptions: (1) the images, which are close to one another in the visual feature space, would have similar emotion distributions in the categorical emotion space; (2) the distribution of a test image can be approximately modeled as a linear combination of the distributions of the training images. Specifically, the combination coefficients are learned in the feature space and transferred to the emotion distribution space. The method is also extended to a more general setting, where multiple features are available. The optimal weights for each feature are automatically learned to reflect the importance of different features.

One intuitive method using deep architecture is to replace the cross-entropy loss for classification with some distribution-based losses, such as KL divergence (Yang et al. 2017):

$$\mathcal{L}_{KL} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_i^j \ln \hat{y}_i^j, \tag{13}$$

where y_i^j and $\hat{y}_i^j$ are the ground truth and predicted probability of the jth emotion category for image $\mathbf{x}_i$. The joint classification and distribution learning (JCDL)

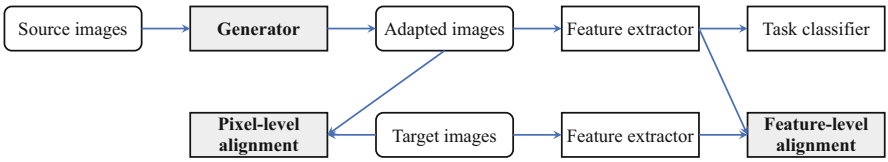

Fig. 6 A generalized domain adaptation framework for IEA with one labeled source domain and one unlabeled target domain. The gray-scale rectangles represent different alignment strategies. Most existing domain adaptation methods can be obtained by employing different component details, enforcing some constraints, or slightly changing the architecture

models both emotion classification and distribution learning simultaneously (Yang et al. 2017).

6.5 Domain Adaptation

Domain adaptation aims to learn a transferable model from a labeled source domain that can perform well on another sparsely labeled or unlabeled target domain (Zhao et al. 2021). Most recent methods focused on the unsupervised setting with a two-stream deep architecture: one stream for training a task model on the labeled source domain, and the other stream for aligning the source and target domains, as shown in Fig. 6. The main difference of existing domain adaptation methods lies in the alignment strategy, which includes discrepancy-based, adversarial discriminative, adversarial generative, and self-supervision-based methods (Zhao et al. 2021).

CycleEmotionGAN++ (CEGAN++) (Zhao et al. 2021) is one state-of-the-art domain adaptation method for IEA. CEGAN++ aligns the source and target domains at both pixel-level and feature-level. First, an adapted domain is generated to perform pixel-level alignment by improving CycleGAN (Zhu et al. 2017) with a multi-scale structured cycle-consistency loss. Dynamic emotional semantic consistency (DESC) is enforced to preserve the emotion labels of the source images during image translation. Second, feature-level alignment is conducted when learning the task classifier. The final objective loss is the combination of task loss, mixed CycleGAN loss, and DESC loss.

7 Released Datasets

In this section, we introduce some datasets that are widely used for performance evaluation of IEA. For clarity, we organize these datasets based on different emotion labels and IEA tasks, i. e., average dimensional values, dominant emotion category, probability distribution, and personalized emotion labels.

Average Dimensional Values *The International Affective Picture System* (*IAPS*) (Lang et al. 1997) is an emotion evoking image set in psychology with 1182 documentary-style natural color images depicting complex scenes, such as portraits, babies, animals, landscapes, etc. Each image is associated with an empirically derived mean and standard deviation (STD) of VAD ratings in a 9-point rating scale by about 100 college students (predominantly US-American). *The Nencki Affective Picture System* (*NAPS*) (Marchewka et al. 2014) is composed of 1356 realistic, high-quality photographs with five categories, i. e., people, faces, animals, objects, and landscapes. 204 mostly European participants labeled these images in a 9-point bipolar semantic sliding scale on the VA and approach-avoidance dimensions. *The Emotions in Context Database* (*EMOTIC*) (Kosti et al. 2017) consists of 18,316 images about people in context in non-controlled environments. There are two kinds of emotion labels: 26 emotion categories and the continuous 10-scale VAD dimensions.

Dominant Emotion Category *IAPSa* (Mikels et al. 2005) is subset of IAPS, which includes 246 images. *Abstract dataset* (*Abstract*) contains 228 peer rated abstract paintings without contextual content (Machajdik & Hanbury 2010). *ArtPhoto* is an artistic dataset with 806 art photos obtained from a photo sharing site (Machajdik & Hanbury 2010). The *IAPSa*, *Abstract*, and *ArtPhoto* datasets are categorized into eight discrete categories (Mikels et al. 2005): amusement, anger, awe, contentment, disgust, excitement, fear, and sadness. The relationship between emotion categories and dimensional VA values is summarized in Fig. 7a. *The Geneva affective picture database* (*GAPED*) consists of 520 negative (133 spiders, 158 snakes, 105 human concerns, and 124 animal mistreatment) images, 121 positive (human and animal babies and nature sceneries) images and 89 neutral (inanimate objects) images (Dan-

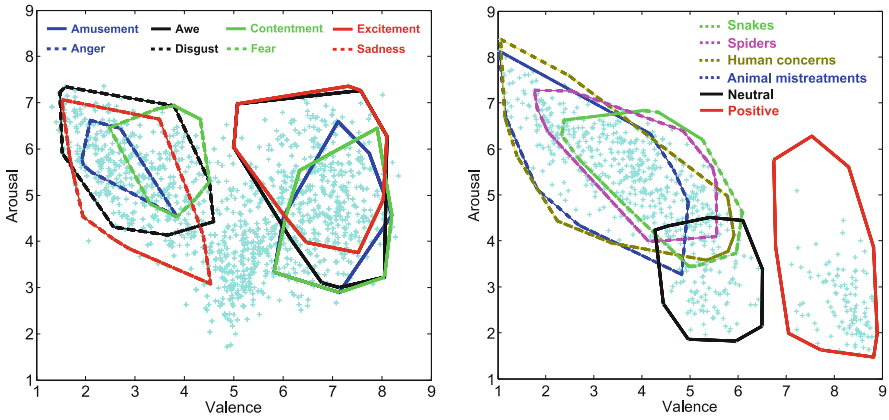

Fig. 7 Representation of the outcome ratings in the valence/arousal space of the (**a**) IAPS and (**b**) GAPED datasets. Polygons represent the surfaces occupied by all the images in a given category

Glauser & Scherer 2011). Besides, these images are also rated with valence and arousal values, ranging from 0 to 100 points. The valence and arousal ratings (changed from [0, 100] to [1, 9]) are shown in Fig. 7b. *Twitter I* (You et al. 2015) consists of 1269 images annotated by 5 Amazon Mechanical Turk (AMT) workers. There are three subsets, i. e., "Five agree" (Twitter I-5), "At least four agree" (Twitter I-4), and "At least three agree" (Twitter I-3). "Five agree" indicates that all the 5 AMT workers labeled the same sentiment label to an image. There are 882 "Five agree" images and all the images receive at least three same votes. *Twitter II* includes 470 positive tweets and 133 negative tweets (Borth et al. 2013) crawled from PeopleBrowsr with 21 hashtags. *EMOd* (Fan et al. 2018) consists of 1019 emotional images with eye-tracking data and different kinds of labels, such as object contour and emotions. *FI* (You et al. 2016) is a large-scale image emotion dataset with 23,308 images labeled using Mikel's emotion categories. The images are obtained by searching from Flickr and Instagram with the eight emotions as keywords and removing noisy data.

Probability Distribution The *Flickr_LDL* and *Twitter_LDL* datasets are constructed to study emotion ambiguity (Yang et al. 2017). There are 10,700 images and 10,045 images in these two datasets, which are labeled by 11 and 8 participants based on Mikel's emotion categories, respectively. Based on the detailed annotations, we can easily obtain the discrete probability distribution of different emotion categories.

Personalized Emotion Labels *Image-Emotion-Social-Net* (*IESN*) (Zhao et al. 2018) is constructed to study personalized emotions. There are more than one million images crawled from Flickr uploaded by 11,347 users. For each image, both the expected emotion from the uploader and actual emotion from each viewer are provided in terms of binary sentiment, Mikel's emotion categories, and continuous VAD values.

8 Experimental Results and Analysis

To give readers a clear understanding of the capabilities of current computational IEA methods, we conduct a series of experiments on different IEA tasks. In this section, we first introduce the evaluation criteria and then report the performance comparison of different representative methods.

8.1 Evaluation Criteria

For emotion classification, the most widely used metric is classification accuracy, which measures the percentage of correctly classified images over all test images (She et al. 2020). For emotion regression, we can use mean squared

error, mean absolute error, and the coefficient of determination to evaluate the results (Zhao et al. 2019). For emotion distribution learning, we can either use the sum of squared difference to measure the performance from the aspect of regression (Zhao et al. 2020), or use distance or similarity metrics (e. g., KL divergence, Bhattacharyya coefficient, Chebyshev distance, Clark distance, Canberra metric, cosine coefficient, and intersection similarity) between two distributions to measure whether the predicted distribution and the ground truth is similar (Yang et al. 2017; Zhao et al. 2020). For image retrieval, there are several evaluation metrics: nearest neighbor rate, first tier, second tier, precision-recall curve, F1 score, discounted cumulative gain (DCG), and average normalized modified retrieval rank (ANMRR) (Zhao et al. 2014; Yao et al. 2020).

We employ accuracy for emotion classification, mean squared error (MSE) for emotion regression, ANMRR for retrieval, and KL divergence for distribution learning. For accuracy, the larger the better; while for MSE, ANMRR, and KL divergence, smaller values indicate better results.

8.2 Supervised Learning Results

For emotion classification and regression, we compare the following methods:

- Traditional methods: principles-of-art based emotion features (PAEF) (Zhao et al. 2014), adjective noun pairs (ANP) with SentiBank (Borth et al. 2013), pretrained AlexNet (Krizhevsky et al. 2012), VGG-16 (Simonyan & Zisserman 2015), and ResNet-101 (He 2016). Support vector machine (SVM) or regression (SVR) with a radial basis function (RBF) kernel is used as the learning model.
- Deep methods: fine-tuned (FT) AlexNet, VGG-16, and ResNet-101, Mldr-Net (Rao et al. 2020), SentiNet-A (Song et al. 2018), WSCNet (She et al. 2020), and PDANet (Zhao et al. 2019).

For emotion retrieval, we compare the performance of the following methods: SIFT (Lowe 1999), HOG (Dalal & Triggs 2005), SentiBank (Borth et al. 2013), Multi-graph learning (MGL) (Zhao et al. 2014), JCDL (Yang et al. 2017), and APSE (Yao et al. 2020).

For emotion distribution learning, the compared methods include: Bayes, SVM, kNN, BP, IIS, BFGS, CPNN (Geng et al. 2013), BCPNN, ACPNN (Yang et al. 2017), CNNR (Peng et al. 2015), DLDL (Gao et al. 2017), and JCDL (Yang et al. 2017).

The results of the above compared methods on emotion classification, regression, retrieval, and distribution learning are shown in Fig. 8. From these results, we can conclude that:

1. Traditional hand-crafted low-level features in computer vision, such as SIFT and HOG, do not perform well on IEA tasks. For example, in Fig. 8c, the retrieval performance of SentiBank is much better than SIFT and HOG on the IAPSa dataset.

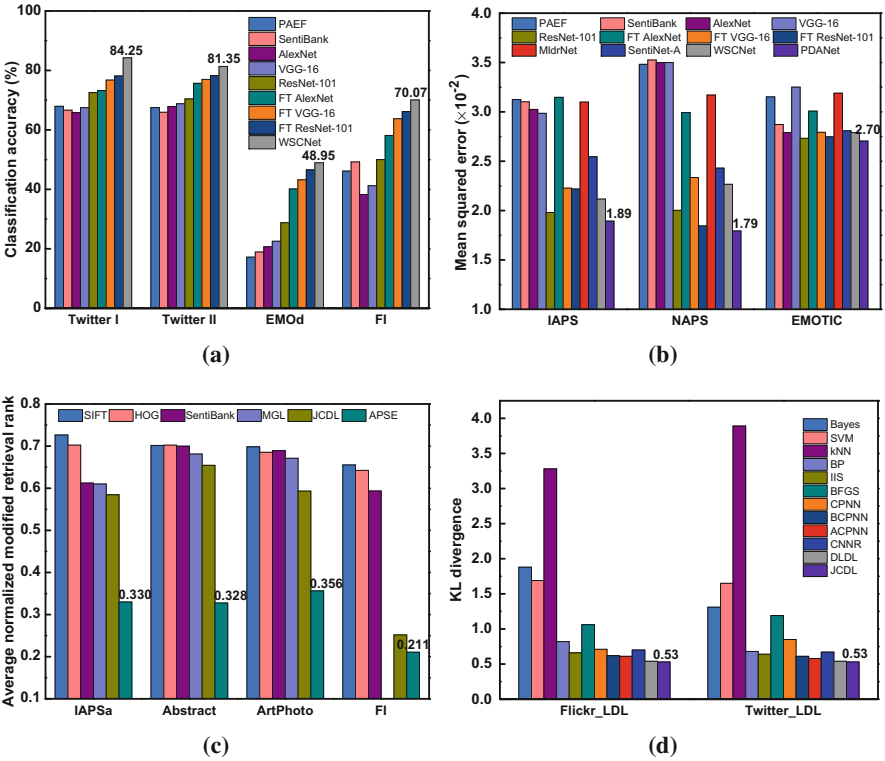

Fig. 8 Performance comparison of supervised learning methods for different IEA tasks, i. e., emotion classification, regression, retrieval, and distribution learning. (**a**) Classification. (**b**) Regression. (**c**) Retrieval. (**d**) Distribution learning

2. Pretraind CNN features, especially the ones extracted from deep models (e. g., ResNet-101), achieve comparable and even better results as compared to hand-crafted specific features, such as PAEF and SentiBank, which demonstrates the generalization ability of deep features to new applications. For example, in Fig. 8a, the pretrained ResNet-101 features achieve 4.63% and 5.92% performance gains on the Twitter I dataset for emotion classification as compared to PAEF and SentiBank.
3. Generally, fine-tuned deep models perform better than pretrained models. This is reasonable, since the pretrained models do not consider the specific characteristics of emotion-related features, while fine-tuned deep models can learn to adapt to the emotion datasets.
4. Deeper models usually perform better, which can be clearly observed when comparing AlexNet and ResNet-101 in Fig. 8a and b.
5. Specially designed models perform the best, such as APSE in Fig. 8c and PDANet in Fig. 8b; by modeling the specific characteristics of emotion, such as polarity-emotion hierarchy and attention mechanisms, these method can better bridge the affective gap.

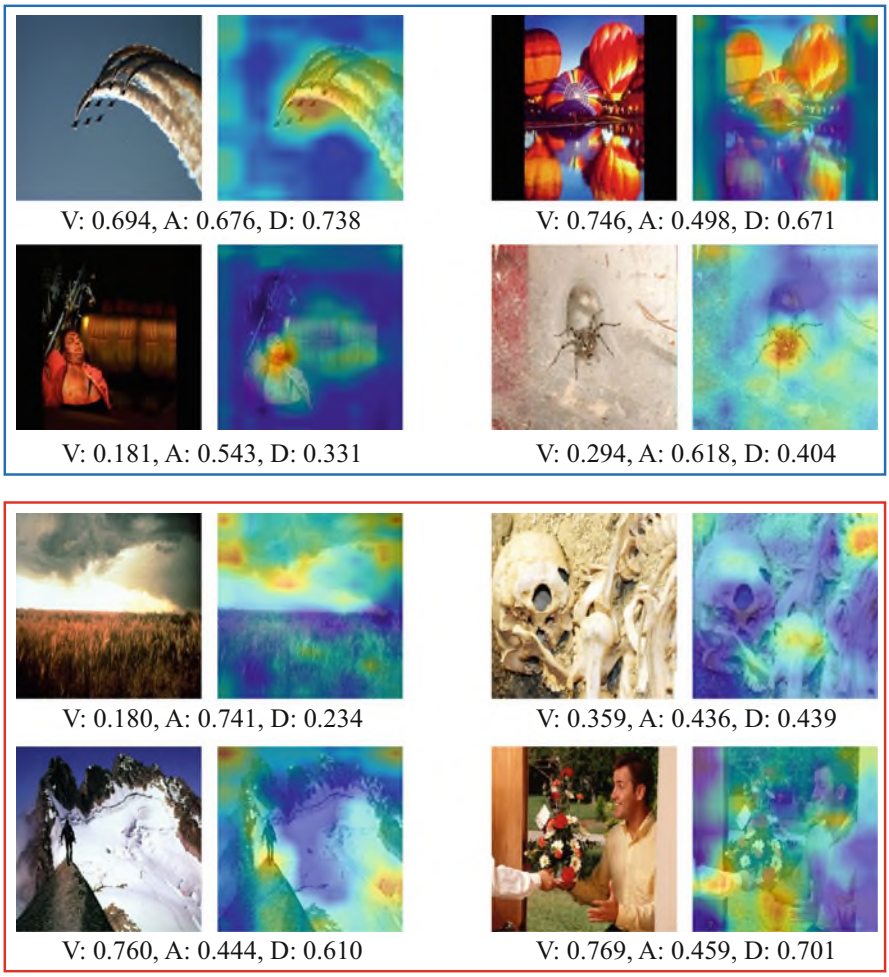

Fig. 9 Visualization of the learned attention maps by PDANet Zhao et al. (2019). From left to right in each image pair are: original image from the test set and the combination of image and heat map. The ground truth VAD values are shown below each pair. Red regions indicate more attention. The attention in the above four examples in the blue rectangle can focus on the salient and discriminative regions, while the below in the red rectangle are failure cases

We visualize the learned attention of PDANet (Zhao et al. 2019) using the heat map generated by the Grad-Cam algorithm (Selvaraju et al. 2017) to show the model's interpretability. The results are shown in Fig. 9. More results on other visualizations can be found in our papers (Yang et al. 2017; Zhao et al. 2019; She et al. 2020; Yao et al. 2020). From the above four examples in the blue rectangle, we can see that PDANet can successfully focus on the salient and discriminative regions that determine the emotion of the whole image. For example, in the top right corner,

the attention learned by PDANet focuses on the colorful balloons, which is strongly related to the positive emotion. We also show some failure cases in the red rectangle. As can be seen, for these cases, the background and foreground are difficult to be distinguished or the background is complex.

8.3 Domain Adaptation Results

For unsupervised domain adaptation for IEA, we report the performance comparison between CycleEmotionGAN++ (CEGAN++) with the following baselines:

- Source-only: directly transferring the model trained on the source domain to the target domain;
- Color style transfer methods: *CycleGAN* (Zhu et al. 2017);
- UDA methods: *ADDA* (Tzeng et al. 2017), *SimGAN* (Shrivastava et al. 2017), and *CyCADA* (Hoffman et al. 2018);
- Oracle: training and testing on the target domain, which can be viewed as an upper bound.

The task classifiers use the ResNet-101 (He 2016) architecture pretrained on ImageNet. Please see (Zhao et al. 2021) for more implementation details. The performance comparisons between CEGAN++ and the above-mentioned approaches are shown in Fig. 10. From the results, we can observe that:

1. Because of the influence of domain shift, directly transferring the models trained on the source domain to the target domain does not perform well. For example, when adapting from ArtPhoto to FI, i. e., training on ArtPhoto and directly testing on FI, the classification accuracy is only 23.86%. The model's low

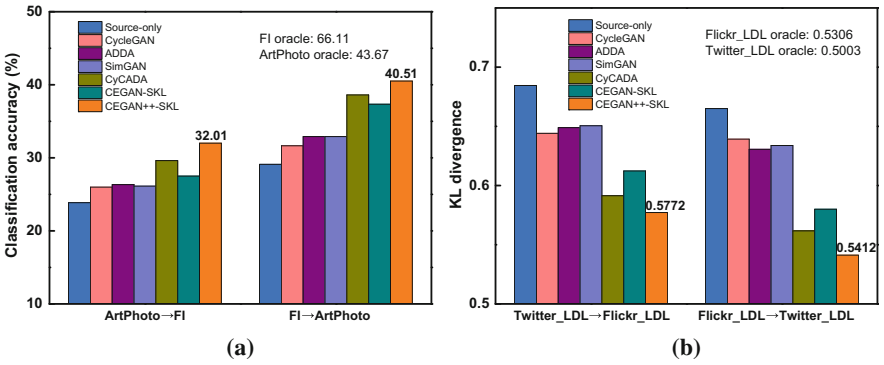

Fig. 10 Domain adaptation results for both emotion classification and distribution learning. For fair comparison and better visualization, the oracle results are shown in detailed numbers in the top right corner. (**a**) Domain adaptation for classification. (**b**) Domain adaptation for distribution learning

transferability from one domain to another motivates the necessity of domain adaptation research.
2. CEGAN++ achieves the best result among all domain adaptation methods for both emotion classification and distribution learning. The superiority of CEGAN++ for adapting image emotions benefits from the following aspects: pixel-level and feature-level alignments to align the source and target domains, dynamic emotional semantic consistency to dynamically preserve the emotion information before and after image translation.
3. There is still an obvious gap between all the domain adaptation methods and the oracle setting that is trained on the target domain. For example, the oracle accuracy on FI is 66.11%, and the best adaptation result is 32.01%. Future efforts are still needed to further bridge the domain shift between different domains.

Figure 11 shows some predicted emotion distributions by different domain adaptation methods on the Twitter-LDL dataset, including one successful example and one failure case. More visualization results can be found in Zhao et al. (2021). From the above example, we can clearly see that the predicted emotion distribution by CEGAN++ is close to the ground truth distribution, which demonstrates its effectiveness for visual emotion adaptation. In the below failure case, we can see that even the oracle does not perform well, which indicates the challenges of IEA, requiring further research efforts.

9 Conclusions and Future Research Directions

We introduced recent advances on image emotion analysis (IEA) from different aspects with the focus on our recent efforts. First, we summarized related psychological studies to understand how emotion is measured. Second, based on the emotion representation models, we defined the key computational problems and widely used supervised frameworks, and then we introduced three major challenges in IEA. Third, we summarized and compared representative methods on emotion feature extraction and learning methods for different IEA tasks. Finally, we briefly described existing datasets and presented an experiment with some of the current state-of-the-art approaches.

Although much research attention has been paid to IEA with promising methods proposed, the overall performance is still not perfect and there is still no solution commonly accepted to address these problems. Many issues in IEA are still open and deserve our further research efforts. We do believe with the progress of multiple disciplines, such as psychology, brain science, and machine learning, IEA will continue to be a hot research topic. At the end, we provide some topics that are well worth considering and investigating.

Context-Aware Image Emotion Analysis Besides extracting discriminative visual features, incorporating available context information can also contribute to the IEA task (Kosti et al. 2020). (1) *Image context*. Similar image content in

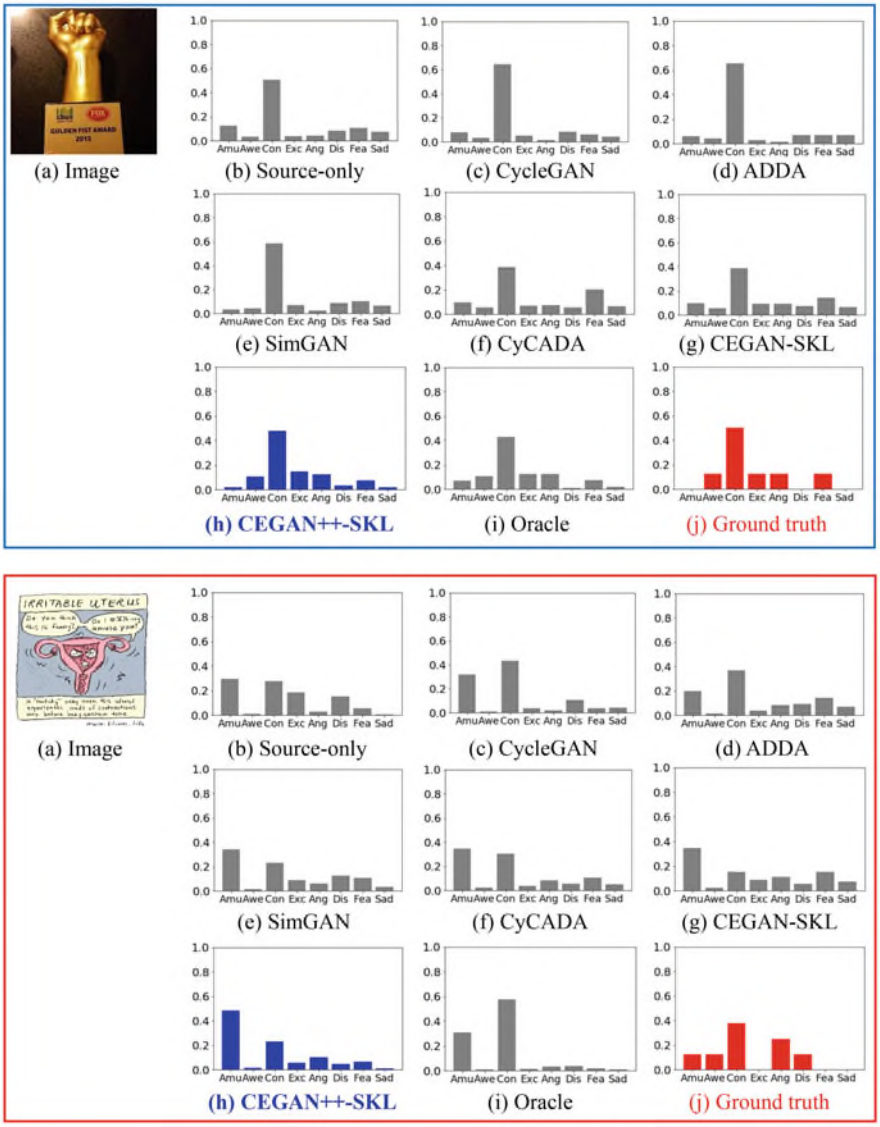

Fig. 11 Visualization of predicted emotion distributions on the Twitter-LDL dataset by CycleEmotionGAN++-SKL (CEGAN++-SKL) Zhao et al. (2021) and several other baselines. In the above example, CEGAN++-SKL can predict similar emotion distribution to the ground truth; while the below example shows a failure case

different contexts might induce totally different emotions, either within an image or across modalities. For example, if we see some soldiers smiling surrounded by flowers, we may feel moved for their contributions to the nation, such as epidemic fighting; but if there is a nearby dead child, we may feel angry for their atrocity.

If we see a famous football player crying on his knees, the audience might feel sad; but if this is after winning a game, the audience especially the team's amateurs my feel excited. (2) *Viewer context*. The context in which a viewer is watching an image and the viewers' prior knowledge (e. g., personality, gender, and culture background) can also contribute a lot to the emotion perception. For example, a viewer's current emotion might be strongly correlated with his/her recent past emotions (Zhao et al. 2018). (3) *Image-viewer interaction*. Humans' emotion perception is a complex process involving both the stimulus and the physical and psychological changes. Combining such implicit and explicit channels are helpful in the final IEA performances.

Determining Intrinsic Emotion Features and Localizing Image Emotions to Image Regions As shown in (Zhao et al. 2014), the emotions of different kinds of images are determined by different features. If we can firstly know the image type, we can select corresponding features that are discriminative for IEA. But what image types should we define for emotion prediction is still unclear. Attempting large scale data-driven approaches is worth trying. Although deep learning based methods achieve promising results for IEA, the explainability on why these methods work, i. e., what features they focus on, has not been fully investigated. Determining the intrinsic features to understand what makes an image amusing, sad or frightening still remains an open problem.

Sometimes, the emotion of an image is determined by the overall appearance of the image. Occasionally, the emotion is reflected by some key image regions. It would be helpful for us to localize these key regions, which can be changed or replaced to change the image emotions (Peng et al. 2014). We can use traditional segmentation methods to segment images into regions and recognize the emotions of each region. Or we can train classifiers to detect the key regions. For example, ANP classifiers are trained hierarchically to localize objects (Chen et al. 2014). More recent emotional region localization methods are based on attention (Zhao et al. 2019) and sentiment maps (She et al. 2020). Besides an emotion classification branch, WSCNet trains another weakly-supervised detection branch to learn the sentiment specific soft map by a fully convolutional network with the cross spatial pooling strategy (She et al. 2020). PDANet jointly considers the spatial and channel-wise attention through which we can obtain the attentive and discriminative regions (Zhao et al. 2019). Jointly combining the advantages of traditional object detection methods and the characteristics of image emotions might motivate new solutions.

Understanding Emotions of 3D Data Most existing works on emotion and sentiment analysis of general images are based on 2D images. But with the wide popularity and public use of somatosensory equipment such as Kinect, more and more 3D data (e. g., 2D images and depth) are created and shared just like personal photos and web videos. Compared with traditional intensity and color images, 3D data contain more information and have several advantages, such as being useful in low light levels and being color and texture invariant (Shotton et al. 2011). Some research efforts have been dedicated to recognizing 3D facial expressions (Sandbach

et al. 2012). However, few works on generalized 3D emotion analysis have been published. To the best of our knowledge, no public emotion dataset of general 3D data is released. Building a large scale 3D emotion dataset is an urgent need and of great value. Using social network data may help to reduce the time-consuming and tedious labelling task. With the rapid development of 3D content analysis, understanding the emotions of 3D data will become a hot research topic.

Image Emotion Analysis in the Wild Existing IEA methods are mainly based on specific settings, such as training on small datasets with limited annotators. However, in real-world applications, the IEA problems are much more complex and difficult. For example, the given datasets might contain inaccurate annotations and much noise that is unrelated to emotion; training data is given incrementally and the emotion categories are becoming more fine-grained gradually; the labeled data is unbalanced across different emotion categories; the test set has different styles from the training set; only limited computing resource is available. How to design an effective and efficient IEA model that can still work under these practical settings is still open.

Novel and Real-world Applications Based on IEA Due to the relatively limited progress in the early years, e. g., low performance, emotion has not been widely deployed in real applications. With recent development of deep learning and large-scale datasets, the IEA performance has been and will continue to be significantly boosted. Therefore, we foresee an emotional intelligence era in the near future with many novel and real-world IEA-based applications. For example, we can understand how artists express emotions through their artworks and use the learned principles in painting education. In fashion advertisement, we can design the best matching between clothes and models to attract users' attention and improve user experience, which can lead to increasing sales.

Security, Privacy, and Ethics of IEA As discussed above, viewers' prior knowledge, such as identity, age, and gender, can contribute to the IEA performance. However, this information is confidential, which should not be shared or leaked. Therefore, protecting the security and privacy must be taken into account in real applications. Further, there is no related law regarding the IEA tasks, especially for personalized scenarios. People might not want their emotion to be recognized and used. From the perspective of ethics, it is important to consider such impact, which requires the joint efforts from different communities, such as psychology, cognitive sciences, and computer science.

Acknowledgments This work is supported by the National Natural Science Foundation of China (Nos. 61701273, 61925107, U1936202, 61876094, U1933114), the National Key Research and Development Program of China Grant (No. 2018AAA0100403), the Natural Science Foundation of Tianjin, China (Nos.20JCJQJC00020, 18JCYBJC15400, 18ZXZNGX00110).

References

Alarcão, S. M., & Fonseca, M. J. (2018). Identifying emotions in images from valence and arousal ratings. *Multimedia Tools and Applications, 77*(13), 17413–17435.

Alarcao, S. M., & Fonseca, M. J. (2019). Emotions recognition using EEG signals: A survey. *IEEE Transactions on Affective Computing, 10*(3), 374–393.

Borth, D., Ji, R., Chen, T., Breuel, T., & Chang, S. F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM International Conference on Multimedia* (pp. 223–232).

Chen, M., Zhang, L., & Allebach, J. P. (2015). Learning deep features for image emotion classification. In *IEEE International Conference on Image Processing* (pp. 4491–4495).

Chen, T., Yu, F. X., Chen, J., Cui, Y., Chen, Y. Y., & Chang, S. F. (2014). Object-based visual sentiment concept analysis and application. In *ACM International Conference on Multimedia* (pp. 367–376).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 886–893).

Dan-Glauser, E.S., & Scherer, K. R. (2011). The Geneva affective picture database (GAPED): A new 730-picture database focusing on valence and normative significance. *Behavior Research Methods, 43*(2), 468–477.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6*(3–4), 169–200.

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). Liblinear: A library for large linear classification. *Journal of Machine Learning Research, 9*, 1871–1874 (2008).

Fan, S., Shen, Z., Jiang, M., Koenig, B. L., Xu, J., Kankanhalli, M. S., & Zhao, Q. (2018). Emotional attention: A study of image sentiment and visual attention. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7521–7531).

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.

Geng, X., Yin, C., & Zhou, Z. H. (2013). Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(10), 2401–2412.

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys, 49*(2), 28.

Guntuku, S. C., Preotiuc-Pietro, D., Eichstaedt, J. C., & Ungar, L. H. (2019). What twitter profile and posted images reveal about depression and anxiety. In *AAAI Conference on Artificial Intelligence* (pp. 236–246).

Hanjalic, A. (2006). Extracting moods from pictures and sounds: Towards truly personalized TV. *IEEE Signal Processing Magazine, 23*(2), 90–100.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., & Darrell, T. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning* (pp. 1989–1998).

Juslin, P. N., & Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research, 33*(3), 217–238.

Kosti, R., Alvarez, J., Recasens, A., & Lapedriza, A. (2020). Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(11), 2755–2766.

Kosti, R., Alvarez, J.M., Recasens, A., & Lapedriza, A. (2017). Emotion recognition in context. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1667–1675).

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).

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). International affective picture system (IAPS): Technical manual and affective ratings. NIMH Center for the Study of Emotion and Attention (pp. 39–58)

Lee, J., & Park, E. (2011). Fuzzy similarity-based emotional classification of color images. *IEEE Transactions on Multimedia, 13*(5), 1031–1039.

Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing* (2020). https://doi.org/10.1109/TAFFC.2020.2981446

Liu, Y., Zhang, D., Lu, G., & Ma, W. Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition, 40*(1), 262–282.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision* (pp. 1150–1157).

Lu, X., Suryanarayan, P., Adams Jr., R. B., Li, J., Newman, M. G., & Wang, J. Z. (2012). On shape and the computability of emotions. In *ACM International Conference on Multimedia* (pp. 229–238).

Machajdik, J., & Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia* (pp. 83–92).

Marchewka, A., Żurawski, Ł., Jednoróg, K., & Grabowska, A. (2014). The Nencki affective picture system (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior Research Methods, 46*(2), 596–610.

Mikels, J. A., Fredrickson, B. L., Larkin, G. R., Lindberg, C. M., Maglio, S. J., & Reuter-Lorenz, P. A. (2005). Emotional category data on images from the international affective picture system. *Behavior Research Methods, 37*(4), 626–630.

Minsky, M. (1986). *The Society of mind*. Simon and Schuster.

Munezero, M. D., Montero, C. S., Sutinen, E., & Pajunen, J. (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing, 5*(2), 101–111.

Pan, S., Lee, J., & Tsai, H. (2014). Travel photos: Motivations, image dimensions, and affective qualities of places. *Tourism Management, 40*, 59–69.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Information Retrieval, 2*(1–2), 1–135.

Parrott, W.G. (2001). *Emotions in social psychology: Essential readings*. Psychology Press.

Peng, K. C., Chen, T., Sadovnik, A., & Gallagher, A. C. (2015). A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 860–868).

Peng, K. C., Karlsson, K., Chen, T., Zhang, D. Q., & Yu, H. (2014). A framework of changing image emotion using emotion prediction. In *IEEE International Conference on Image Processing* (pp. 4637–4641).

Picard, R. W. (2000). *Affective computing*. MIT Press.

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*. Harpercollins College Division.

Plutchik, R., & Kellerman, H. (2013). *Theories of emotion* (Vol. 1). Academic Press.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion, 37*, 98–125.

Rao, T., Xu, M., & Xu, D. (2020). Learning multi-level deep representations for image emotion classification. *Neural Processing Letters, 51*(3), 2043–2061.

Sandbach, G., Zafeiriou, S., Pantic, M., & Rueckert, D. (2012). Recognition of 3d facial expression dynamics. *Image and Vision Computing, 30*(10), 762–773.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review, 61*(2), 81.

Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM, 61*(5), 90–99.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision* (pp. 618–626).

She, D., Yang, J., Cheng, M. M., Lai, Y. K., Rosin, P. L., & Wang, L. (2020). Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection. *IEEE Transactions on Multimedia, 22*(5), 1358–1371.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., & Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1297–1304).

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2107–2116).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Smeulders, A. W., Worring, M., Santini, S., Gupta, A., & Jain, R. (2020). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(12), 1349–1380.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing, 65*, 3–14.

Song, K., Yao, T., Ling, Q., & Mei, T. (2018). Boosting image sentiment analysis with visual attention. *Neurocomputing, 312*, 218–228.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Weblogs and Social Media* (Vol. 10, pp. 178–185).

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7167–7176).

Wang, S., & Ji, Q. (2015). Video affective content analysis: a survey of state-of-the-art methods. *IEEE Transactions on Affective Computing, 6*(4), 410–430.

Wang, W., & He, Q. (2008). A survey on emotional semantic image retrieval. In *IEEE International Conference on Image Processing* (pp. 117–120).

Wei, Z., Zhang, J., Lin, Z., Lee, J. Y., Balasubramanian, N., Hoai, M., & Samaras, D. (2020). Learning visual emotion representations from web data. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 13106–13115).

Xu, C., Cetintas, S., Lee, K., & Li, L. (2014). Visual sentiment prediction with deep convolutional neural networks. arXiv:1411.5731.

Yang, J., She, D., Lai, Y., & Yang, M. H. (2018). Retrieving and classifying affective images via deep metric learning. In *AAAI Conference on Artificial Intelligence* (pp. 491–498).

Yang, J., She, D., & Sun, M. (2017). Joint image emotion classification and distribution learning via deep convolutional neural network. In *International Joint Conference on Artificial Intelligence*, (pp. 3266–3272).

Yang, J., Sun, M., & Sun, X. (2017). Learning visual sentiment distributions via augmented conditional probability neural network. In *AAAI Conference on Artificial Intelligence* (pp. 224–230).

Yang, P., Liu, Q., & Metaxas, D. N. (2010). Exploring facial expressions with compositional features. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2638–2644).

Yang, Y. H., & Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology, 3*(3), 40.

Yanulevskaya, V., Van Gemert, J., Roth, K., Herbold, A., Sebe, N., & Geusebroek, J. (2008). Emotional valence categorization using holistic image features. In *IEEE International Conference on Image Processing* (pp. 101–104).

Yao, X., Zhao, S., Lai, Y. K., She, D., Liang, J., & Yang, J. (2020). Apse: Attention-aware polarity-sensitive embedding for emotion-based image retrieval. *IEEE Transactions on Multimedia* (2020). https://doi.org/10.1109/TMM.2020.3042664

You, Q., Jin, H., & Luo, J. (2017). Visual sentiment analysis by attending on local image regions. In *AAAI Conference on Artificial Intelligence* (pp. 231–237).

You, Q., Luo, J., Jin, H., & Yang, J. (2015). Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI Conference on Artificial Intelligence* (pp. 381–388).

You, Q., Luo, J., Jin, H., & Yang, J. (2016). Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI Conference on Artificial Intelligence* (pp. 308–314).

Yuan, J., Mcdonough, S., You, Q., & Luo, J. (2013). Sentribute: Image sentiment analysis from a mid-level perspective. In *International Workshop on Issues of Sentiment Discovery and Opinion Mining* (pp. 1–8).

Zhan, C., She, D., Zhao, S., Cheng, M. M., & Yang, J. (2019). Zero-shot emotion recognition via affective structural embedding. In *IEEE International Conference on Computer Vision* (pp. 1151–1160).

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1253.

Zhao, S., Chen, X., Yue, X., Lin, C., Xu, P., Krishna, R., Yang, J., Ding, G., Sangiovanni-Vincentelli, A. L., & Keutzer, K. (2021). Emotional semantics-preserved and feature-aligned cyclegan for visual emotion adaptation. *IEEE Transactions on Cybernetics* (2021)

Zhao, S., Ding, G., Gao, Y., Zhao, X., Tang, Y., Han, J., Yao, H., & Huang, Q. (2020). Discrete probability distribution prediction of image emotions with shared sparse learning. *IEEE Transactions on Affective Computing, 11*(4), 574–587.

Zhao, S., Ding, G., Huang, Q., Chua, T. S., Schuller, B. W., & Keutzer, K. (2018). Affective image content analysis: A comprehensive survey. In *International Joint Conferences on Artificial Intelligence* (pp. 5534–5541).

Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T. S., & Sun, X. (2014). Exploring principles-of-art features for image emotion recognition. In *ACM International Conference on Multimedia* (pp. 47–56).

Zhao, S., Jia, Z., Chen, H., Li, L., Ding, G., & Keutzer, K. (2019). Pdanet: Polarity-consistent deep attention network for fine-grained visual emotion regression. In *ACM International Conference on Multimedia* (pp. 192–201).

Zhao, S., Li, Y., Yao, X., Nie, W., Xu, P., Yang, J., & Keutzer, K. (2020). Emotion-based end-to-end matching between image and music in valence-arousal space. In *ACM International Conference on Multimedia* (pp. 2945–2954).

Zhao, S., Ma, Y., Gu, Y., Yang, J., Xing, T., Xu, P., Hu, R., Chai, H., & Keutzer, K. (2020). An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In *AAAI Conference on Artificial Intelligence* (pp. 303–311).

Zhao, S., Wang, S., Soleymani, M., Joshi, D., & Ji, Q. (2019). Affective computing for large-scale heterogeneous multimedia data: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications, 15*(3s), 93.

Zhao, S., Yao, H., Gao, Y., Ding, G., & Chua, T. S. (2018). Predicting personalized image emotion perceptions in social networks. *IEEE Transactions on Affective Computing, 9*(4), 526–540.

Zhao, S., Yao, H., Gao, Y., Ji, R., & Ding, G. (2017). Continuous probability distribution prediction of image emotions via multi-task shared sparse regression. *IEEE Transactions on Multimedia, 19*(3), 632–645 (2017)

Zhao, S., Yao, H., Yang, Y., & Zhang, Y. (2014). Affective image retrieval via multi-graph learning. In *ACM International Conference on Multimedia* (pp. 1025–1028).

Zhao, S., Yue, X., Zhang, S., Li, B., Zhao, H., Wu, B., Krishna, R., Gonzalez, J. E., Sangiovanni-Vincentelli, A. L., Seshia, S. A., & Keutzer, K. (2021). A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems* (2021)

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision* (pp. 2223–2232).

Zhu, X., Li, L., Zhang, W., Rao, T., Xu, M., Huang, Q., & Xu, D. (2017). Dependency exploitation: a unified cnn-rnn approach for visual emotion recognition. In *International Joint Conference on Artificial Intelligence* (pp. 3595–3601).