Affective Image Content Analysis: Two Decades Review and New Perspectives

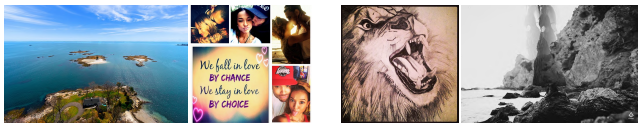
Sicheng Zhao, *Senior Member, IEEE*, Xingxu Yao, Guoli Jia, Jufeng Yang, Guiguang Ding, Tat-Seng Chua, Björn W. Schuller, *Fellow, IEEE*, Kurt Keutzer, *Life Fellow, IEEE*

Abstract—Images can convey rich semantics and induce various emotions in viewers. Recently, with the rapid advancement of emotional intelligence and the explosive growth of visual data, extensive research efforts have been dedicated to affective image content analysis (AICA). In this survey, we will comprehensively review the development of AICA in the recent two decades, especially focusing on the state-of-the-art methods with respect to three main challenges – the affective gap, perception subjectivity, and label noise and absence. We begin with an introduction to the key emotion representation models that have been widely employed in AICA and description of available datasets for performing evaluation with quantitative comparison of label noise and dataset bias. We then summarize and compare the representative approaches on (1) emotion feature extraction, including both handcrafted and deep features, (2) learning methods on dominant emotion recognition, personalized emotion prediction, emotion distribution learning, and learning from noisy data or few labels, and (3) AICA based applications. Finally, we discuss some challenges and promising research directions in the future, such as image content and context understanding, group emotion clustering, and viewer-image interaction.

Index Terms—Affective computing, image emotion, emotion feature extraction, machine learning, emotional intelligence

1 Introduction

In the book "The Society of Mind" [1], Minsky (a Turing Award winner in 1970) claimed that "*The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without emotions.*" Although emotions play a vitally important role in machine and artificial intelligence, much less attention has been paid to affective computing than objective semantic understanding, such as object classification in computer vision. The rapid development of artificial intelligence has made remarkable success in semantic understanding and raised higher demand to emotional interaction. For example, the companion robots that can recognize and express emotions can provide more harmonious companionship for human beings, especially the elderly and single children. To have human-like emotions, machines should first understand how humans express emotions through multiple channels, such as speech, gesture, facial expression, and physiological signals [2]. While other signals can be easily suppressed or masked, physiological signals that are controlled by the sympathetic nervous systems are independent of humans' will and thus provide more reliable information. However, to capture accurate physiological signals is quite difficult and impractical, as it requires special types of wearable sensors. On the other hand, the recent convenient access of cameras in mobile devices and wide popularity of social networks (such as Twitter, Flickr, and Weibo) have enabled people to habitually share their experiences and express their opinions online using images and videos together with text [3]. Recognizing the affective content of this large volume of multimedia data provides an alternate way to understand users' behaviors and emotions.



(a) Positive (b) Negative

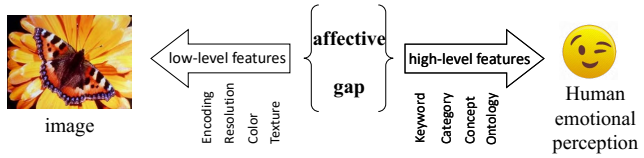
Fig. 1. Examples of relevance and importance of AICA to infer humans' emotional status. Images are from the FI dataset [4].

As we know, "a picture is worth a thousand words", which indicates that images can convey rich semantics. Different from existing research on analyzing the cognitive aspects of images, such as object detection and semantic segmentation, affective image content analysis (AICA) focuses on understanding the semantics at a higher level – the affective level, *i.e.* understanding the emotions that can be induced by the images in viewers, which is more challenging. The automatic inference of humans' emotional status using AICA can help to evaluate their psychological health, discover affective anomaly, and prevent extreme behaviors to themselves and even to the whole society. For example, in Fig. 1, the users posting images (b) are more likely to have negative emotions and take revenge on society to express their dissatisfaction than the users posting images (a).

1.1 Main Goals and Challenges

Main Goals. Given an input image, AICA mainly aims to (1) recognize the emotions that can be induced to specific viewers or to the majority (Based on psychology, the emotions might be

- *S. Zhao and G. Ding are with BNRist, Tsinghua University, Beijing 100084, China. (e-mail: schzhao@gmail.com, dinggg@tsinghua.edu.cn).*
- *X. Yao, G. Jia, and J. Yang (corresponding author) are with the College of Computer Science, Nankai University, China (e-mail: yxx_hbgd@163.com, exped1230@gmail.com, yangjufeng@nankai.edu.cn).*
- *T.-S. Chua is with the School of Computing, National University of Singapore, Singapore (e-mail: dcscts@nus.edu.sg).*
- *B. W. Schuller is with the Department of Computing, Imperial College London, UK (e-mail: bjoern.schuller@imperial.ac.uk).*
- *K. Keutzer is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA (e-mail: keutzer@berkeley.edu).*



(a) Affective gap overview

(b) Affective gap examples

Fig. 2. Illustration of the affective gap. (a) Overview: the commonly extracted low-level features cannot well represent high-level emotions. (b) Examples: the first pair of images have a similar object (rose) but evoke different emotions, while the second pair of images exhibit entirely different content (car versus house) but evoke similar emotions.

(a) Within images (b) Across modalities

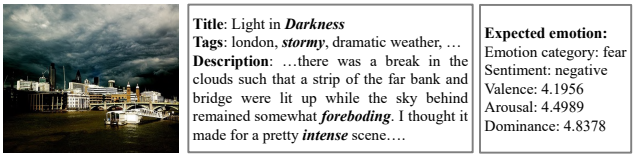
Fig. 3. The context information also plays an important role in AICA. (a) The image without and with the detailed scene context evoke different emotions (surprise vs. happy). (b) The textual contexts can also influence the emotion perception of the same image (sad vs. excited).

represented in different models, *e.g.* categorical or dimensional. Please see Section 2 for details.), (2) analyze what stimuli contained in the image evoke such emotion (*e.g.* specific objects or color combinations), and (3) apply the recognized emotions to different real-world applications to improve the ability of emotional intelligence.

Challenges. (1) Affective Gap. Similar to the semantic gap in computer vision, the affective gap is one main challenge for AICA, which can be defined as "the lack of coincidence between the features and the expected affective state in which the user is brought by perceiving the signal" [5], as shown in Fig. 2.

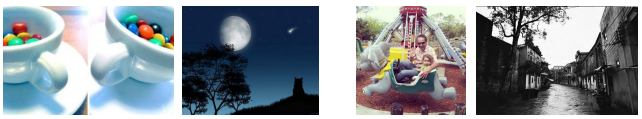
To bridge the affective gap, researchers primarily focus on extracting discriminative features that can better distinguish the difference among different emotions, ranging from hand-crafted features like Gabor [6], Gist [7], artistic elements [8], artistic principles [9], and adjective noun pairs (ANPs) [10] to deep ones like convolutional neural networks (CNNs) [11, 4] and regions [12]. Based on the assumption that different viewers can reach a consensus on the perceived emotions of images, these AICA methods mainly assign an image with the dominant (average) emotion category (DEC). This task can be performed as a traditional single-label learning problem.

Besides extracting visual features, incorporating available context information can also contribute to the AICA task [13], as shown in Fig. 3. The same image under different contexts may evoke different emotions. For example, in Fig. 3 (a), if we just see the kid, we may feel surprise based on his expression; but with the context that the kid is blowing the candles to celebrate his birthday, it is more likely to make us feel happy. In Fig. 3 (b), if we see a volleyball player crying, we may feel sad; but if there is a comment for the image, "Finally, we won, after ten years!", we, especially the volleyball amateurs of the team, may feel excited.



Comments to this image from different viewers	Personalized emotion labels
Wow, that is ***fantastic***...it looks so ***incredible***, like a painting...slightly unreal. That sky is ***amazing***.	Emotion: awe, Sentiment: positive V: 7.121 A: 4.479 D: 6.635
Yup a fave for me as well. ***Exciting*** drama at its best.	Emotion: excitement, Sentiment: positive V: 7.950 A: 6.950 D: 7.210
Hey, it really ***frightened*** me! My little daughter just looked ***scared***.	Emotion: fear, Sentiment: negative V: 2.625 A: 5.805 D: 3.625

Fig. 4. Illustration of the perception subjectivity [17]. For the original image (a) uploaded to Flickr, different viewers may have different emotion perceptions (b). The emotion labels are obtained using the keywords in italic based on the comments from these viewers.



(a) Domain shift

	train on ArtPhoto	train on FI
test on ArtPhoto	44.87	**29.49**
test on FI	**20.17**	66.81

(b) Performance evaluation

Fig. 5. Illustration of domain shift. (a) The images from ArtPhoto [8] and FI [4] datasets have different styles: artistic vs. social. (b) The emotion classification performance (%) significantly drops if the trained dataset is different from the tested dataset on both ArtPhoto and FI datasets by fine-tuning the ResNet-101 model [18].

(2) Perception Subjectivity. Different viewers may have totally different emotional reactions to the same image, which is caused by many personal and contextual factors, such as the cultural background, personality and social context [14, 15, 16]. For example, for the "Light in Darkness" image in Fig. 4(a), viewers who are interested in capturing natural phenomena are probably excited to see this spectacle, while the viewers who are scared of thunder and storm might feel fear. This fact causes the so-called subjective perception problem. Therefore, for this highly subjective variable, simply predicting the DEC is insufficient, since it cannot well reflect the difference among different viewers.

To tackle the subjectivity issue, we can conduct two kinds of AICA tasks [15]: for each viewer, we can predict personalized emotion perceptions; for each image, we can assign multiple emotion labels. For the latter one, we can employ multi-label learning methods, which associate one image with multiple emotion labels. However, since the importance or extent of different emotion labels is in fact unequal, emotion distribution learning would make more sense, which aims to learn the degree to which each emotion describes the image [16].

(3) Label Noise and Absence. Recent methods on AICA based on deep learning, especially CNN, have achieved promising results. However, training these models requires large-scale labeled data, which is prohibitively expensive and time-consuming to obtain, not only because labeling the emotions in ground-truth generation

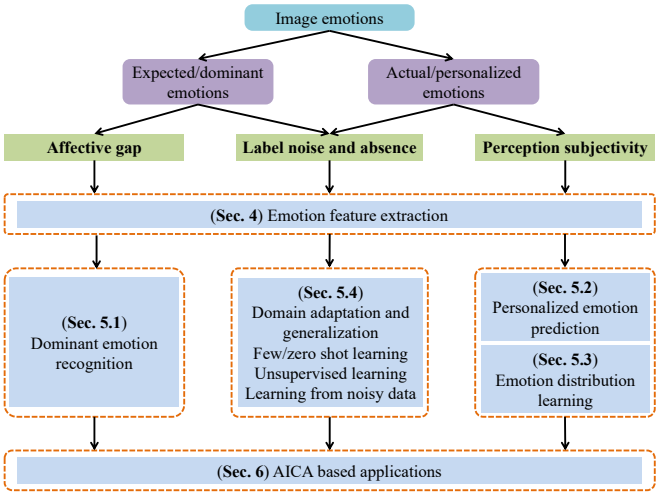


Fig. 6. Organization of different reviewed technical components.

is highly inconsistent, but also because in some cases like artistic works only experts are able to provide reliable labels. In real-world applications, there might be only few or even no labeled emotion data. How to deal with this situation is significantly worth investigating. Unsupervised/weakly supervised learning and few/zero shot learning are two interesting directions.

One possible solution is to leverage the unlimited amount of web images with associated tags as labels [19]. However, such tags can be incomplete and noisy. An image might be associated with tags that are unrelated or remotely related. How to learn from noisily labeled images is the main challenge. Imposing some constraints for visual representation based on the semantic correlations between image and text is one direct solution. Firstly learning text models and embeddings in unsupervised or semi-supervised manners and then denoising the keyword labels can help to "clean" the label noise.

Furthermore, if we have sufficient labeled data in one domain, such as abstract paintings, how can we effectively transfer the well-trained models to another unlabeled or sparsely labeled domain? Because of the presence of *domain shift* or *dataset bias* [20, 21], direct transfer often results in poor performance, as shown in Fig. 5. Specifically, Panda et al. [22] classified the dataset bias in AICA into two categories. One is positive set bias. Due to the lack of diversity in visual concepts for each emotion category (*e.g.* amusement) in the source domain, the models learned based on such data are easily to memorize all its idiosyncrasies and lose the ability to generalize to the target domain. The other is negative set bias. The rest of the dataset (*e.g.* the data coming from other categories excluding amusement) in the source domain does not well represent the rest of the visual world. For example, some of the negative samples from the target domain are confused with the positive samples in the source domain. As a result, the learned classifiers might be overconfident. Domain adaptation and domain generalization might help to address this issue.

1.2 Organization of This Survey

In this survey, we concentrate on reviewing the state-of-the-art methods on AICA and outlining research trends. First, we introduce the brief history in Section 1.3 and its comparison with other related topics in Section 1.4. Second, we describe the widely-used emotion representation models in Section 2. Third, we summarize the available datasets for performing AICA evaluation in Section 3 and quantitatively compare the label noise and dataset bias. Fourth, based on the main goals and challenges in Section 1.1, we summarize and compare the representative approaches on emotion feature extraction, learning methods (for dominant emotion recognition, personalized emotion prediction, emotion distribution learning, and learning from noisy data or few labels), and AICA based applications in Sections 4, 5, and 6, respectively, as shown in Fig. 6. Finally, we discuss potential research directions to pursue in Section 7.

1.3 Brief History

Affective Computing. Before affective computing was known by this term, early first works include a 1978 filed patent on an analyzer for determining the emotion of a speaker speech [23], and scientific papers on generation of affect in synthesized speech in 1990 [24], or the recognition of facial expressions by neural networks in 1992 [25].

Since Minsky proposed the emotion recognition problem of intelligent machines [1], much attention has been paid to emotion related research, such as the definition of emotional intelligence [26]. In 1997, Picard first proposed the concept of affective computing [27]: "affective computing is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena". Some influential events include: the first International Conference on Affective Computing and Intelligent Interaction (ACII) by IEEE and AAAI in 2005, the foundation of the Association for the Advancement of Affective Computing (AAAC) in 2007 (originally named HUMAINE Association), the first ever public 'emotion challenge' held at Interspeech 2009, the launch of the IEEE Transactions on Affective Computing (TAFFC) in 2010, the first International Audio/Visual Emotion Challenge and Workshop (AVEC) in 2011, the proposal of the Emotional and Social Signals in Multimedia area in ACM Multimedia 2014, the first international challenge to also feature physiology (AVEC) in 2015, and the first ACII Asia in 2018, *etc.*

Affective Image Content Analysis. The development of AICA begins in the psychology and behavior research, such as the International Affective Picture System (IAPS) [28, 29], to investigate the relation between visual stimuli and emotion. One of the first emotion recognition methods is based on low-level holistic Wiccest and Gabor features [6]. Since then, several representative hand-crafted features have been designed, such as the low-level artistic elements [8], mid-level artistic principles [9], and high-level Adjective Noun Pairs (ANPs) [10]. In 2014, transfer learning is conducted from a CNN in which parameters are pre-trained by large-scale data [30]. To tackle the subjectivity challenge, both personalized emotion prediction [31, 15] and emotion distribution learning [14, 32, 33, 34] are considered. More recently, domain adaptation [35, 36, 37] and zero-shot learning [38] are studied for the label absence challenge. The representative milestones in both general affective computing and AICA are summarized in Fig. 7.

1.4 Comparison with Other Related Topics

Comparison with Affective Computing of Other Modalities. Affective content analysis has also been widely studied in other modalities, such as text [39, 40], speech acoustics [41, 42] and linguistics [43], music [44, 45], sound [46], facial expression [47, 48, 49, 50], video [51, 52], physiological signals [53, 54, 55],

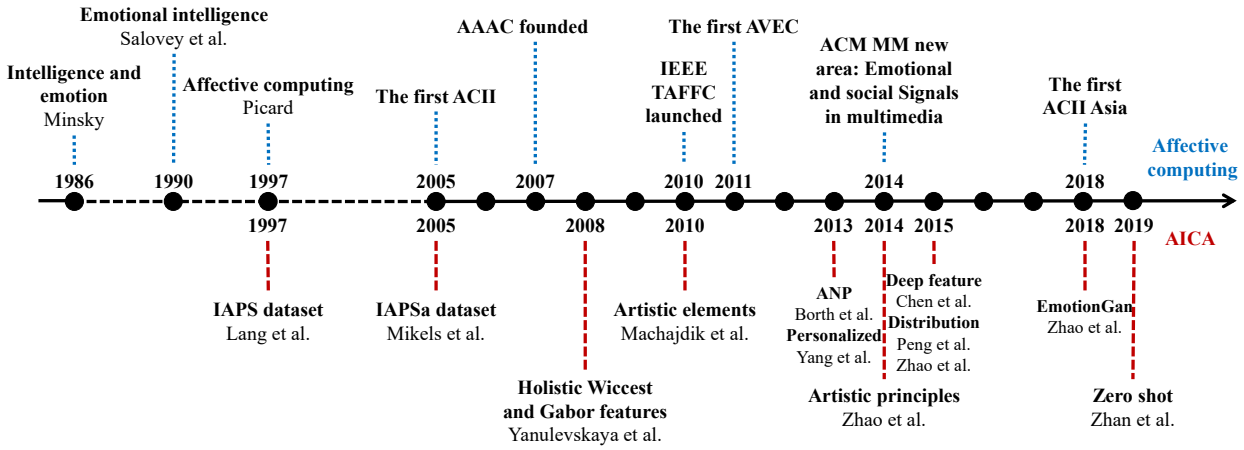


Fig. 7. Milestones in both general affective computing (above line, blue) and affective image content analysis (below line, red).

and multi-modal data [56, 57, 58, 59]. Although the employed emotion models and learning methods are similar, there is a clear difference between affective computing of images and other modalities, especially the extracted features to represent emotions. While the surveys on other modalities are well-conducted, there is no comprehensive survey on AICA. As an image is an important channel to express emotions, we believe an in-depth analysis of AICA could boost the development of affective computing. One preliminary version on this survey was previously introduced in our IJCAI 2018 conference paper [60]. As compared to the conference paper, this journal version has the following five aspects of extensive enhancements. First, the detailed challenges and a brief history are incorporated. Second, we summarize and compare more representative works on emotion models, available datasets, emotion features, and learning methods. Third, we conduct extensive experiments to fairly compare the effectiveness of different AICA methods. Fourth, we add some AICA-based applications. Finally, we discuss more potential research directions.

Comparison with Computer Vision. The task of AICA is often composed of three steps: human annotation, visual feature extraction, and learning of mapping between visual features and perceived emotions [61]. Although the three steps seem to be very similar to computer vision (CV, the third step is a mapping learning between visual features and image labels, such as an object), there are significant differences between AICA and CV.Take object classification and emotion classification for instance. (1) Even if the semantic gap is bridged in object classification, there still exists an affective gap. For example, an image with a lovely dog and an image with a barking dog can evoke different emotions. (2) Object is an objective concept (both a lovely dog and a barking dog are dogs), while emotion is a relatively subjective concept (happy and fear for the two images). (3) Correspondingly, object classification belongs to the cognitive aspects of images, while AICA focuses on the affective level. Object classification is mainly studied by the CV community, while AICA is an interdisciplinary task requiring psychology, cognitive science, multimedia, and machine learning, *etc.*

2 Emotion Models from Psychology

In psychology, there are several different affective computing related concepts, such as emotion, affect, mood, and sentiment. Discussing the difference or correlation of these concepts is out of the scope of this survey. Interested readers can refer to [67] for more details. Appraisal theory is well-known for explaining the development of emotional experience [68]. It accounts for individual variability in emotional reactions to the same stimulus. According to the Ortony, Clore and Collins (OCC) model [69], the emergence of emotions originates from the cognitive evaluation or appraisal of stimuli in terms of events, agents, and objects. How individuals actually perceive and interpret the stimuli determines how emotions might emerge.

TABLE 1
Representative emotion models employed in AICA.

Model	Ref	Type	Emotion states/dimensions
Ekman	[62]	CES	happiness, sadness, anger, disgust, fear, surprise
Mikels	[29]	CES	amusement, anger, awe, contentment, disgust, excitement, fear, sadness
Plutchik	[63]	CES	($\times$ 3 scales) anger, anticipation, disgust, joy, sadness, surprise, fear, trust
Parrott	[64]	CES	a tree hierarchical grouping with primary, secondary and tertiary emotion categories
Sentiment		CES	positive, negative, (and neutral)
VA(D)	[65]	DES	valence-arousal(-dominance)
ATW	[66]	DES	activity-temperature-weight

Psychologists mainly employ two kinds of emotion representation models to measure emotion: categorical emotion states (CES) and dimensional emotion space (DES), as shown in Table 1. CES models classify emotions into a few basic categories. The simplest CES model is binary *positive* and *negative* (polarity) [70, 55]. In such cases, "emotion" is often called "sentiment", which sometimes also includes *neutral*. Since sentiment is too coarse-grained, some relatively fine-grained emotion models are designed, such as Ekman's six emotions (anger, disgust, fear, happiness, sadness, surprise) [62] and Mikels's eight emotions (amusement, anger, awe, contentment, disgust, excitement, fear, and sadness) [29]. With the development of psychological theories, categorical emotions are becoming increasingly diverse and fine-grained such as by also considering social emotions [71]. Besides the eight basic emotion categories (anger, anticipation, disgust, fear, joy, sadness, surprise, trust), Plutchik [63] organized each of them into 3 intensities which thus provides a richer set. For example, the 3 intensities of joy and fear are ecstasy$\rightarrow$joy$\rightarrow$serenity and terror$\rightarrow$fear$\rightarrow$apprehension, respectively. Another representative

TABLE 2
Comparison between CES and DES.

	CES	DES
understandability	easy	difficult
describability	limited	unlimited
perspective	qualitative	quantitative
examples	Mikels, Plutchik	VAD
granularity	coarse-grained	fine-grained
AICA task	classification, retrieval	regression, retrieval

CES model is Parrott's tree hierarchical grouping [64], which represents emotions with primary, secondary and tertiary categories. For example, a three-level emotion hierarchy is designed as two basic categories (positive and negative) at level-1, six categories (anger, fear, joy, love, sadness, and surprise) at level-2 and 25 fine-grained emotion categories at level-3.

Although CES models are easy for users to understand, limited emotion categories cannot well reflect the complexity and subtlety of emotions. Further, psychologists have not reached a consensus on how many discrete emotion categories should be included. Differently, DES models employ continuous 2D, 3D, or higher dimensional Cartesian space to represent emotions, such as valence-arousal-dominance (VAD) [65] and potentially added intensity, novelty, or others, and activity-temperature-weight [66]. VAD is the most widely used DES model [72], where valence represents the pleasantness ranging from positive to negative, arousal represents the intensity of emotion ranging from excited to calm, and dominance represents the degree of control ranging from controlled to in control. In practice, dominance is difficult to measure and is often omitted, leading to the commonly used two-dimensional VA space [5]. Theoretically, every emotion can be represented as a coordinate point in the Cartesian space. However, the absolute continuous values are difficult for users to distinguish, which constraints the employment of DES models.

The comparison between CES and DES is shown in Table 2. Compared to CES, DES is able to represent finer-grained and more comprehensive emotions, which reflects their difference on granularity and describability, respectively. Further, instead of being independent from each other, they are actually related. The relationship between CES and DES and the transformation from one to the other are studied in [73, 74]. For example, positive valence relates to a happy state, while negative valence relates to a sad or angry state; a relaxed state relates to low arousal, while anger relates to high arousal. To further distinguish for example anger and fear (both negative valence, but high arousal), one needs dominance (high for anger, but low for fear). CES and DES are mainly employed in classification and regression tasks, respectively, with discrete and continuous emotion labels. As a result, the employed learning models are different. For the affective image retrieval task, both models can be employed with different emotion distance measurements (*e.g.* Mikels' emotion wheel [15] for CES and Euclidean distance for DES). If we discretize DES into several constant values, we can also use it for classification [66]. We can consider easing DEC comprehension difficulties in raters by ranking based labeling.

Another relevant concept worth mentioning is that emotion in response to multimedia can be expected, induced, or perceived emotion. Expected emotion is the emotion that the multimedia 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. We do not aim discussing the difference or correlation of various emotion models in this survey and believe that the achievements from psychology and cognitive science are beneficial for the AICA task.

3 DATASETS

In the early years, the affective datasets only contain small-scale images built from psychology or artistic communities. With the development of digital photography and online social networks, an increasing number of large-scale datasets have been created by crawling the images posted on Internet. We summarize all the datasets for AICA in Table 3.

3.1 Brief Introduction to Different Datasets

The International Affective Picture System (**IAPS**) [28] is an image dataset for visual emotional stimuli used in experimental investigations of emotion and attention in psychology [28]. The dataset contains 1,182 documentary-style natural color images with various contents or scenes, such as portraits, babies, animals, landscapes, *etc.* About one hundred college students took part in the VAD rating on a 9-point scale. The mean and standard deviation (STD) of scores for each image can be derived easily.

The subset A of IAPS (**IAPSa**) [29] is collected from IAPS to characterize the images by a descriptive discrete emotion category. Specifically, 203 negative images and 187 positive images are selected, and then labeled by twenty undergraduate participants. To the best of our knowledge, it is the first affective image dataset which is labeled using a discrete emotion category.

The **Abstract** dataset [8] consists of 280 paintings which are combinations only of color and texture. They are annotated by about 230 people, and each image is voted 14 times. For each image, the category obtaining the most number of votes is regarded as the ground truth. After filtering the images whose votes are inconclusive, 228 images are retained.

The **Artphoto** dataset [8] contains 806 artistic photos collected from an art sharing site. The photos are obtained by searching the site with the emotion categories as keywords. The ground truth of each image is determined by the user who uploads it.

The Geneva affective picture database (**GAPED**) [75] contains 730 pictures which are collected to make full use of visual emotion stimuli. Several specific types of negative or positive content are presented in these images. The 520 negative images, 121 positive images, and 89 neutral images are labeled by 60 people ranging from 19 to 43 years (mean=24, STD=5.9). In addition, the continuous VA scales are rated from 0 to 100 points.

The **MART** dataset [76] contains 500 abstract paintings collected from more than 20,000 artworks of professional artists guided by an art historian. There are 25 participants (11 females and 14 males) annotating these images with negative or positive rating. Each person annotated 145 paintings on average. The **devArt** [76] is a dataset of Amateur paintings from the deviantArt (dA) website. The 500 paintings created by 406 different authors are labeled by 60 people including 27 females and 33 males.

The **Twitter I** dataset [77] consists of 1,269 images. A total of 5 Amazon Mechanical Turk (AMT) workers were employed to label the images. The dataset contains three subsets, including "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 reached an agreement on the sentiment label

TABLE 3
Released datasets for AICA, where '# Images' and '# Annotators' represent the total number of images and annotators (f: female, m: male).

Dataset	Ref	# Images	Type	# Annotators	Emotion model	Label detail
IAPS	[28]	1,182	natural	≈100 (half f)	VAD	empirically derived mean and standard deviation
IAPSa	[29]	390	natural	20 (10f,10m)	Mikels	at least one emotion category for each image
Abstract	[8]	280	abstract	≈230	Mikels	the detailed votes of all emotions for each image
ArtPhoto	[8]	806	artistic	–	Mikels	one DEC for each image
GAPED	[75]	730	natural	60	Sentiment, VA	one DEC and average VA values for each image
MART	[76]	500	abstract	25 (11f,14m)	Sentiment	one DEC for each image
devArt	[76]	500	abstract	60 (27f,33m)	Sentiment	one DEC for each image
Twitter I	[77]	1,269	social	5 per image	Sentiment	one sentiment category for each image
Twitter II	[10]	603	social	3 per image	Sentiment	one sentiment category for each image
VSO	[10]	≈500,000	social	–	Plutchik	one emotion category for each image
MVSO	[78]	7.36M	social	–	Plutchik	one emotion category for each image
Flickr I	[79]	354,192	social	6,735	Ekman	one emotion category for each image
Flickr II	[80]	60,745	social	3 per image	Sentiment	one sentiment category for each image
Instagram	[80]	42,856	social	3 per image	Sentiment	one sentiment category for each image
Emotion6	[14]	1,980	social	432	Ekman+neutral, VA	the discrete probability distribution
FI	[4]	23,308	social	225	Mikels	one DEC for each image
IESN	[15]	1,012,901	social	118,035	Mikels, VAD	the emotion of involved users for each image
T4SA	[81]	1,473,394	social	-	Sentiment+neutral	one sentiment category for each image
B-T4SA	[81]	470,586	social	-	Sentiment+neutral	one sentiment category for each image
Comics	[82]	11,821	comic	10 (5f,5m)	Mikels	one DEC for each image
Event	[83]	8,748	social	3 each image	Sentiment+neutral	one sentiment category for each image
EMOTIC	[84]	18,316	social	3 each image	Ekman, VAD	one DEC and VAD values for each image
EMOd	[85]	1,019	natural	3	Sentiment+neutral	object contour, object name, sentiment category
WEBEmo	[22]	268,000	social	-	Parrott	one DEC for each image
LUCFER	[86]	3.6M	social	-	Plutchik, VAD, context	one DEC, average VAD values, and context for each image
FlickrLDL	[16]	10,700	social	11	Mikels	the discrete probability distribution
TwitterLDL	[16]	10,045	social	8	Mikels	the discrete probability distribution

of an image. Twitter I-5 contains 882 images, while all the images obtain at least three same votes on sentiment.

The **Twitter II** dataset [10] includes 470 positive images and 133 negative images collected using over 20 twitter hashtags. Three different labeling runs, namely image-based, text-based, and image-text based, were conducted by 3 random AMT workers (each worker for each run), respectively. The final selected images all receive unanimous sentiment votes.

The images labeled with 1,553 ANPs in **VSO** [10] are retrieved and downloaded using the Flickr API. The corresponding ANP should be contained in the title, tag, or caption of the image. As psychological principles for construction of datasets, Plutchik's Wheel of Emotions covers 3 intensities based on 8 basic emotions. **MVSO** [78] is the extension of the VSO. The dataset consists of more than 7.36M images annotated with ANPs from 12 different languages including Arabic, Chinese, Dutch, English, French, German, Italian, Persian, Polish, Russian, Spanish, and Turkish. A total of 4,342 English ANPs were constructed.

Flickr I [79] is proposed to study the correlation between emotions and friends' discussions on the images. It contains 354,192 images posted by 4,807 users, of which all the comments and tags are included. To model the friends' interactions well, the detailed information of users is also recorded in the dataset, including ID, alias, and contact list.

Flickr II [80] and **Instagram** [80] are collected from Flickr and Instagram, respectively, based on query keywords. The sentiment polarity labels are provided via online crowdsourcing. Specifically, each image was shown to three random workers, who should choose a rating from the five scales including highly positive, positive, neutral, negative, and highly negative. The final ground-truth of each image is determined by the major ratings of polarity. After discarding the images that are labeled as 'neutral' or received opposite sentiment annotations, 48,139 positive and 12,606 negative images are left in the Flickr II dataset, while Instagram contains 33,076 positive and 9,780 negative images.

Emotion6 [14] contains 1,980 images which are obtained from Flickr by six category keywords and corresponding synonyms. Each image is annotated by 15 participants with both valence-arousal scores and discrete emotion distribution. The categories include Ekman's six basic emotions [62] and neutral.

FI [4] is a large-scale affective image dataset constructed based on Mikel's emotions. All the images are collected from Flickr and Instagram with the eight emotions as search keywords. A total millions of weakly labeled images are crawled. After deleting noisy data, a total of 225 AMT workers were employed to assess the emotions of images. Finally, 23,308 images receive at least three agreements from the assigned annotators. The dataset is widely used in the field of AICA.

The **IESN** dataset [15] consists of more than one million images crawled from Flickr uploaded by 11,347 users, and it is constructed to study the personalized emotion perception. Therefore, various metadata of corresponding images are also collected, including tags, descriptions, comments, and uploaders' social context. For each image, the labels of expected emotion from the uploader and actual emotion from the viewer are both generated. In addition, by leveraging the VAD norms of 13,915 English lemmas [87], the average values of VAD are computed as label of DES. According to the descriptions and the comments of the images, the emotion distribution is also easy to obtain.

The **T4SA** dataset [81] consists of about one million tweets and corresponding images. According to the textural sentiment classification, the images are classified into positive, negative, and neutral. However, the dataset contains 501,037 positive, 214,462 negative, and 757,895 neutral images, which is very imbalanced. As a balanced subset, **B-T4SA** [81] is extracted from T4SA, in which there is an equivalent of 156,862 images in each class.

The **Comics** dataset [82] is composed of 11,821 images selected from seventy comics, including One Piece, Spiderman, Sponge Bob, The Avengers, *etc*. A total of 10 participants (mean age=20.3 years) were employed to label these images using Mikel's eight emotion categories. The dataset is further divided into two subsets: A Comics subset and a Manga subset. The

Dataset	ρ_{-1}	ρ_{+1}	ρ_m
FI	0.169	0.082	0.108
Flickr_LDL	0.278	0.098	0.129
Twitter_LDL	0.404	0.068	0.097
IAPSa	0.315	0.257	0.283
Abstract	0.387	0.257	0.307
Artphoto	0.245	0.271	0.257
Comics	0.171	0.189	0.180
GAPED	0.170	0.227	0.187
Event	0.148	0.056	0.074
Twitter I	0.290	0.156	0.209
Twitter II	0.492	0.198	0.264
Flickr	0.184	0.176	0.180
Instagram	0.195	0.179	0.187
WEBEmo	0.098	0.115	0.107

(a) Label noise

	FI	Flickr_LDL	Twitter_LDL	IAPSa	Abstract	Artphoto	Comics	GAPED	Event	Twitter I	Twitter II	Flickr	Instagram	WEBEmo
FI	0.875	0.729	0.673	0.818	0.739	0.764	0.609	0.636	0.752	0.650	0.647	0.722	0.690	0.602
Flickr_LDL	0.754	0.882	0.905	0.701	0.587	0.621	0.582	0.566	0.787	0.610	0.824	0.555	0.556	0.448
Twitter_LDL	0.732	0.842	0.918	0.675	0.609	0.596	0.542	0.404	0.793	0.634	0.832	0.519	0.523	0.444
IAPSa	0.734	0.759	0.830	0.818	0.674	0.584	0.519	0.727	0.762	0.543	0.740	0.549	0.551	0.464
Abstract	0.644	0.591	0.646	0.571	0.739	0.665	0.476	0.647	0.738	0.622	0.639	0.586	0.573	0.528
Artphoto	0.677	0.469	0.584	0.740	0.739	0.807	0.575	0.747	0.717	0.571	0.613	0.702	0.628	0.495
Comics	0.706	0.721	0.814	0.766	0.522	0.627	0.813	0.546	0.707	0.614	0.739	0.562	0.601	0.465
GAPED	0.605	0.616	0.589	0.662	0.609	0.571	0.584	0.919	0.624	0.575	0.630	0.585	0.585	0.472
Event	0.694	0.753	0.841	0.532	0.674	0.503	0.490	0.515	0.931	0.795	0.849	0.587	0.556	0.452
Twitter I	0.735	0.803	0.867	0.662	0.652	0.534	0.494	0.505	0.882	0.823	0.815	0.616	0.582	0.455
Twitter II	0.700	0.816	0.908	0.532	0.630	0.491	0.470	0.333	0.826	0.654	0.824	0.505	0.498	0.437
Flickr	0.769	0.657	0.664	0.727	0.652	0.795	0.595	0.677	0.842	0.780	0.647	0.804	0.731	0.501
Instagram	0.777	0.708	0.747	0.779	0.696	0.745	0.592	0.727	0.853	0.803	0.748	0.786	0.784	0.495
WEBEmo	0.700	0.573	0.563	0.740	0.609	0.683	0.601	0.747	0.750	0.713	0.681	0.677	0.669	0.804

(b) Dataset bias

Fig. 8. Quantitative comparison and ranking of different datasets. (a) Estimated label noise ratio of different datasets. ρ_{-1} means the noise ratio of negative sentiment, and ρ_{+1} means the noise ratio of positive sentiment. ρ_m is the mean noise ratio. (b) The confusion matrix between different datasets on sentiment polarity classification, which can reflect the bias between any two datasets.

Comics subset includes samples from European and American comics which are drawn in the realism style, while the Manga subset contains Asian comics with an abstract style.

The **Event** dataset [83] has 8,748 images that are obtained from Microsoft Bing by using search keywords from 24 event categories. The types of events are diverse, including personal and actual public events. In the annotation process, an image could be retained if at least 2 out of 3 crowdworkers reach an agreement on its label. The sentiment labels of these event images contain positive, negative, and neutral.

The **EMOTIC** dataset [84] contains 18,316 images which are selected from the MSCOCO [88] and Ade20k [89] datasets and were downloaded via the Google search engine based on 26 emotional keywords. The dataset has two types of annotated information. One is the DES comprising 26 emotions, while the other is continuous VAD dimensions. A total of 23,788 people (66% males, 34% females) are annotated in the images.

The **EMOd** dataset [85] consists of 1,019 emotional images, in which 321 images are selected from IAPS [28] and 698 images are collected via online image search engine. Each image owns eye-tracking data collected from 16 participants, and is labeled with detailed information including object contour, object sentiment, semantic category, and high-level perceptual attributes such as image aesthetics and emotions. Three undergraduate students were employed to label the characteristic of the objects in the images. The emotion of the object is labeled by 'neutral' when the agreements are different.

The **WEBEmo** dataset [22] is a large-scale web emotion dataset constructed based on Parrott's hierarchical emotion model. First, about 300,000 weakly labeled images are collected by searching with keywords for each of the 25 emotions. Then, the duplicate images and those with non-English tags are removed, leading to 268,000 high-quality images.

The **LUCFER** dataset [86] contains over 3.6M images, which are labeled with 24 emotional categories from Plutchik's model, 3 continuous emotional dimensions, and image contexts. By combining contexts and emotions, a total of 275 emotion-context pairs are generated. First, 80,649 images are collected from the *wild*, and 35,239 images of them pass the AMT workers' validation. With over 30 thousand images, a large-scale dataset is obtained employing *Bing*'s feature available in its *Image Search API*.

Flickr_LDL [16] and **Twitter_LDL** [16] are constructed to study the emotion ambiguity. Each image is labeled by more than one viewer based on their own emotional reactions. Flickr_LDL contains 10,700 images extracted from VSO dataset. A total of 11 participants were hired to view each image and label the images with one of the Mikel's eight emotions. Twitter_LDL is collected from Twitter with various sentiment keywords, and 8 viewers were employed to label these images within the same eight emotions. Finally, 10,045 images are retained after deduplication.

3.2 Comparison Among Different Datasets

Here we compare several released datasets from the perspectives of label noise and dataset bias for readers to better understand how to select required datasets in real applications.

3.2.1 Label Noise

For quantitative comparison in terms of label noise, we estimate the noise rate by pre-training CNN with softmax loss [101]. It is assumed that the probability of positive (+1) images being assigned to negative (-1) is $\rho_{+1} = p(\hat{y} = -1|y = +1)$, where y represent the ground truth and $\hat{y}$ is the predicted label. Similarly, the probability of negative images being assigned to positive is $\rho_{-1} = p(\hat{y} = +1|y = -1)$. If there is no label noise, the values of ρ_{-1} and ρ_{+1} will approach 0. In Fig. 8 (a), we estimate the values of ρ_{+1} and ρ_{-1} using the algorithm in [102, 101]. First, we resort images based on the predicted probability values from small to large:

$$\hat{p}(\hat{y}|x_{n+1}) \geq \hat{p}(\hat{y}|x_n), \hat{p}(\hat{y}|x_{N_{\hat{y}}}) = \max \hat{p}(\hat{y}|x_n), \quad (1)$$

where $n = 1, 2, \cdots, N_{\hat{y}}$, and $N_{\hat{y}}$ denotes the number of images that is predicted as $\hat{y}$. Second, we build a fitting Gaussian function $g(\cdot)$ between the predicted probability and the corresponding numbers of it:

$$n_{\hat{p}(\hat{y}_n|x_n)} = g_{\hat{y}_n}(\hat{p}(\hat{y}_n|x_n)), N_{\hat{p}(\hat{y}_n|x_n)} = \int_0^1 g_{\hat{y}_n}(\hat{p}(\hat{y}_n|x_n)), \quad (2)$$

where $n_{\hat{p}(\hat{y}_n|x_n)}$ is the number of images that are predicted as $\hat{y}_n$ with probability $\hat{p}(\hat{y}_n|x_n)$. Third, the noise ratio is regarded as the deviation between the probability that obtains the maximum value on the fitting function and 1:

$$\begin{aligned} \rho_{+1} &= 1 - \arg\max g_{+1}(\hat{p}(\hat{y}_n = +1|x_n)), \\ \rho_{-1} &= 1 - \arg\max g_{-1}(\hat{p}(\hat{y}_n = -1|x_n)), \end{aligned} \quad (3)$$

Finally, we compute the mean noise ratio ρ_m weighted by the proportion of each category:

$$\rho_m = \sum\nolimits_i^c v_i \cdot \rho_i, \quad (4)$$

TABLE 4
Summary of hand-crafted features on different levels that have been used for AICA. '# Feat' indicates the dimension of each feature.

Feature	Ref	Level	Short description	# Feat
WLDLV	[90]	low	orientation and length information of lines	12
EFS	[91]	low	luminance-warm-cool fuzzy histogram, saturation-warm-cool fuzzy histogram, luminance contrast	10, 7, 2
Eleven Groups	[92]	low	shape, edge, texture, polynomial, image statistics	691
LOW_C	[7]	low	Gist, HOG2x2, self-similarity and geometric context color histogram features	17,032
Elements	[8]	low	color: mean saturation, brightness and hue, emotional coordinates, colorfulness, color names, Itten contrast, Wang's semantic descriptions of colors, area statistics; texture: Tamura, Wavelet and gray-level co-occurrence matrix	97
MPEG-7	[66]	low	color: layout, structure, scalable color, dominant color; texture: edge histogram, texture browsing	≈200
Shape	[93]	low	line segments, continuous lines, angles, curves	219
IttenColor	[94]	low	color co-occurrence features and patch-based color-combination features	16,485
Attributes	[7]	mid	scene attributes	102
Sentributes	[95]	mid	scene attributes, eigenfaces	109
Constructs	[96]	mid	roundness, angularity, complexity	3
Composition	[8]	mid	level of detail, low depth of field, dynamics, rule of thirds	45
Aesthetics	[97]	mid	figure-ground relationship, color pattern, shape, composition	13
Principles	[9]	mid	principles-of-art: balance, contrast, harmony, variety, gradation, movement	165
SIFT	[98]	mid	bag-of-visual-words on SIFT, latent topics	330
FS	[8]	high	number of faces and skin pixels, size of the biggest face, amount of skin w.r.t. the size of faces	4
ANP	[10]	high	semantic concepts based on adjective noun pairs	1,200
Expressions	[99]	high	automatically assessed facial expressions (anger, contempt, disgust, fear, happiness, sadness, surprise, neutral)	8
HLCs	[100]	high	object information and scene information	1,205

where v_i is the proportion of images from the i-th category. ρ_i is the noise ratio of the i-th category. c is the number of categories. From the table, we can observe that Event is the dataset with the least label noise. However, Abstract has the largest label noise. It is mainly because the abstract images are difficult to distinctly understand, leading to more label noise.

3.2.2 Dataset Bias

To explore the dataset bias of emotion recognition, we conduct extensive transfer learning experiments among various datasets by training a classifier on one dataset and testing on another. These datasets have wide ranges on the number of images, ranging from a few hundred to a few hundred thousand. Meanwhile, these datasets include various types, such as natural images and abstract images. We split each dataset into 80% training images and 20% test images. For a fair experiment, we fine-tune a ResNet-50 on each dataset and test the model on all the datasets.

The confusion matrix is shown in Fig. 8 (b), where each row represents the results of one model on different datasets. We have some observations as follows. First, there is a larger bias between datasets that belong to different types, such as abstract, comic, and natural images. For example, the types of Abstract, Comics, and GAPED are abstract, comic, and natural, respectively. The model trained on Comics only obtains 0.522 and 0.546 accuracy on Abstract and GAPED datasets. Second, the bias between the two datasets is not mutual. This means that the evaluated bias is smaller when the cleaner one of two datasets is regarded as the target. For instance, as the shown experimental results between the Event (the cleanest dataset, ρ_m = 0.074) and other datasets, the results are better when Event is the target dataset. Third, the similarity of class distribution can influence the dataset bias. For example, Twitter II and Flickr_LDL (Twitter_LDL) have similar class distribution, where there are significantly more positive images than negative images. However, Flickr and Instagram have the balanced class distribution, which is different from that of Twitter II. Therefore, it is observed that the model trained on Twitter II has better performances on Flickr_LDL and Twitter_LDL than those on Flickr and Instagram. Finally, as the most widely-used dataset, FI has the best generalization ability among these datasets. As shown in the first row of Fig. 8 (b), the model trained on FI dataset obtains over 60% classification accuracy on all the datasets.

4 Emotion Feature Extraction

To describe image emotion with informative representations, many studies explore extraction of various types of features. In terms of hand-crafted features, we introduce the designed features on different levels. Besides, we review the emerging deep features in recent years with the development of CNNs.

4.1 Hand-crafted Features

The hand-crafted features on different levels focus on different aspects, as summarized in Table 4.

4.1.1 Low-level Features

Various low-level features were designed to represent emotional content in the early years, although they lack reasonable interpretation. As one pioneering study on hand-crafted features, [90] explores the relationship between line directions and image emotion. Specifically, horizontal lines relate to the static horizon and always express the relaxation and calmness, while the direct and clear vertical lines convey eternality and dignity. Lines with various directions can convey different emotions. With the degree of long, thick, and straight lines increasing, the expressed emotions will be stronger. However, to capture more informative information, guided by the theories of color psychology, Wang et al. [91] constructed three kinds of image features in an orthogonal three-dimensional emotion factor space, respectively. The features include luminance-warm-cool representation, saturation-warm-cool-contrast representation, and contrast-sharpness representation. By mining these proposed emotional information, they designed an emotion-based image retrieval system.

When it comes to hand-crafted features, one cannot ignore a milestone [8], in which different types of features are combined. Particularly, color and texture are the representative low-level features in composition. Color is represented with a 70-dimensional vector consisting of eight kinds of statistical information, while texture is encoded with a 27-dimensional vector containing three types of image statistics. Later in [66], another complex feature combination consisting of color and texture, named MPEG-7, is proposed. Besides, a fuzzy similarity relation is applied to computing the weights of different components. Zhang et al. [92] listed eleven groups of features referring to texture, shape, and further,

which are extracted from several transforms of the image. The influence of the visual shapes on the image emotions is systematically explored in [93]. The experimental results demonstrate the effectiveness of shape features for emotion prediction. In addition, Gist, 2×2 Histogram of Oriented Gradients (HOG), self-similarity, and geometric context color histogram features are widely used due to their ability to represent the distinct visual scenes [7]. Based on Itten's color wheel, Sartori et al. [94] investigated the different color combinations in abstract paintings, and used the factors to analyze the emotions evoked in the viewers.

4.1.2 Mid-level Features

Compared with low-level features, mid-level ones are easier to understand by humans, and they can largely bridge the gap between low-level features and high-level emotions. Patterson and Hays [7] designed a large-scale attribute database, named the SUN attribute database, which consists of 102 attributes that belong to different types, including materials, surface properties, and others. Based these the mid-level attributes, Yuan et al. [95] proposed an image emotion recognition algorithm, named Sentribute, in which the facial expression detected based eigenface is also added as a crucial element to determine the polarity of emotions.

As an essential feature in artworks, harmonious composition [8] is introduced for emotion representation. Guided by the theories of art, Wang et al. [97] designed more interpretable and understandable features, in which the description of the contrast between the figure and the ground is included. Apart from the studies that extract features of the overall image with the single scale, there are also some researchers that pay attention to mining informative representation in the multi-scale blocks of each image. For instance, Rao et al. [98] used two different image segmentation types for extracting multi-scale blocks in each image. With SIFT as the basic feature, they employed bag-of-visual-words (BoVW) to encode each block, and then adopted probabilistic latent semantic analysis to further estimate the latent topic used as a mid-level representation. This study reveals that the BoVW can well model the emotional information of different local regions, and the features directly extracted from the whole image may lead to wrong classification results.

Based on different artistic principles, the combined emotion representation named principles-of-art is proposed [9], consisting of balance, emphasis, harmony, variety, gradation, and movement. As the milestone of mid-level representation, principles-of-art features obtained the state-of-the-art performance at that time. Without features of hundreds of dimensions, three visual characteristics, including roundness, angularity, and visual complexity, are proposed in [96], each of which is only a one-dimensional scalar. For the three mid-level representations, it has been demonstrated that they can be used to recognize emotions well.

4.1.3 High-level Features

High-level features refer to the semantic information of images, which are easy to understand and can directly evoke emotions in viewers. In [8], Machajdik and Hanbury extracted content that draws the attention of the viewers and has great effects on the emotions, including human faces and skin. Facial expression, an important high-level feature, always plays a decisive role in the process of evoking emotions. It is usually classified following Ekman's 'Big Six' into six basic emotions, which include anger, disgust, fear, happiness, sadness, and surprise. In [99], based on the compositional features of image patches, Yang et al. detected and analyzed the categories of facial expression.

As a milestone, a large visual sentiment ontology named SentiBank is proposed by Borth et al. [10]. It contains 1,200 concepts, and each concept represents an ANP, like *beautiful flower*, which provides powerful semantic representation. First, based on 24 emotions defined in Plutchik's theory, the authors retrieved related adjectives and paired them with frequently used nouns. The final remaining 1,200 ANPs cover 178 adjectives and 313 nouns after filtering the redundant ANPs. ANPs provided a novel solution to bridge the "affective gap" at that time, because they are easy to be mapped into emotions. Later, as the extension of the SentiBank, a large-scale multilingual visual sentiment ontology (MVSO) is proposed in [78]. Particularly, there are 4,342 English ANPs in MVSO. In [100], high-level concepts (HLCs), including objects and places, are introduced to bridge the affective gap between image content and evoked emotion. The HLCs are explicitly derived from pre-trained CNNs, and subsequently, a linear admixture model is employed to capture the relations between emotions and HLCs.

4.2 Deep Features

In recent years, with the rapid development of CNNs, learning-based features have shown superior performance in the field of AICA. In the beginning, each region in an image is treated equally in the learning process, and global features are extracted for different tasks. Later, based on the theories of psychology that emotional content is always involved in some important regions, more and more studies have focused on how to extract informative local features.

4.2.1 Global Features

Based on a deep CNN model, the classifiers of the 1,200 ANP concepts are trained using Caffe. The newly trained deep model named DeepSentiBank [103] performs better than non-deep SentiBank on sentiment prediction. Benefiting from transfer learning, Xu et al. [30] transferred the parameters of a CNN trained on the large-scale dataset (ImageNet) to the task of predicting sentiments. They extracted a 4096-dimensional representation from the fully connected (FC) layer FC7 layer and a 1000-dimensional representation from FC8, respectively, as the image-level features. The experimental results on Twitter II dataset demonstrate that the features from FC7 exhibit an advantage in describing the emotional content, because the activations from the FC7 layer can characterize more aspects of the image than those of FC8 layer.

The progressive CNN (PCNN) [77] is another milestone, which is pre-trained using half-a-million weakly labeled images from SentiBank. In the learning process, the training instances that have large difference between the two polarities are kept for the next round of training. With the iterative training, the noisy data can be removed progressively, such that the trained model is more robust when transferring to those small-scale strongly labeled datasets. In [11], Chen et al. explored two methods of obtaining image features. One method uses off-the-shelf CNN (pre-trained on ImageNet) features to train a one-vs-all linear SVM classifier. The other method is to initialize the parameters of the pre-trained AlexNet and replace the 1,000-way classification layer with 8-way emotion category outputs. Subsequently, the network can be trained end-to-end from raw images to the specific classes.

High-level semantic features may be not enough as emotional representation in some images, especially in abstract paintings. To

capture different types of information in the images, some studies integrate the multi-level deep features generated in CNNs. Rao et al. [104] proposed an end-to-end architecture that consists of three parallel neural networks, including an Alexnet, an aesthetics CNN (A-CNN), and a texture CNN (T-CNN). Before being fed into the network, the images are segmented into different levels of patches. Subsequently, the three sub-networks are exploited to extract deep representations at three-levels, respectively, *i.e.*, image semantics, image aesthetics, and low-level visual features. Zhu et al. [105] extracted the multi-level features from different layers in CNNs. The output from each layer is fed into a bidirectional gated recurrent unit (Bi-GRU) model to exploit the dependency between them. Finally, the features output from the two ends are concatenated as the emotional representations. It is further considered that a Gram matrix can capture powerful texture features [106]; hence, Yang et al. [107] proposed a sentiment representation consisting of elements in Gram matrices from different layers.

4.2.2 Local Features

To emphasize the informative regions that contain attractive emotional content, local features are drawing more and more attention in recent studies. Considering fine-grained details, the features of local patches are extracted at multiple scales in [11]. Next, they are aggregated with the Fisher Vector for a more compact representation. In [108], apart from investigating general emotion content using multiple instance learning, Liu et al. also detected facial expression and emotional objects to constitute the emotion factors when computing visual saliency.

How to find the crucial emotion-related regions based on an image-level label is a question worth exploring. Based on the descriptive visual attributes, You et al. [109] adopted an attention model to discover the local regions that evoke sentiment in viewers, and then extracted the features of them to improve the performance on visual sentiment analysis. Yang et al. [110] utilized an off-the-shelf object detection tool to generate bounding box candidates. After removing the redundant proposals, the selected regions have a high probability of containing an object and accordingly obtain a high sentiment score. Further, the features of selected regions and the holistic images are jointly used for classification. However, the process of selecting proper image regions is very time-consuming in this work. To simplify and improve the step of selecting informative regions, a unified CNN that contains a classification branch and a detection branch is proposed in [12]. In the detection branch, the soft sentiment map is generated by combining all the class-wise feature responses. The comprehensive localized information can be derived by coupling the holistic feature and the sentiment map. Later in [111], both spatial and channel-wise attended features are incorporated into the final representation for visual emotion regression in VAD space. To effectively utilize various information from multiple layers, Rao et al. [112] proposed a multi-level region-based CNN framework to find the emotional response of the local regions. First, the feature pyramid network (FPN) is employed to extract multi-level deep representations. Following this, the regions of interest (ROIs) are detected based on the region proposal method, and their features in multiple levels are concatenated for image emotion classification. The work has achieved the best classification performance on several benchmark datasets up to now. To obtain an informative feature embedding for affective image retrieval, Yao et al. [113] conducted polarity- and emotion-specific attention on the lower layers and higher layers, respectively. The attended features from different layers are integrated by cross-level bilinear pooling to generate the final representation.

TABLE 5
Experimental comparison between local and global features measured by average classification accuracy and rank. 'L' denotes the local features, and 'G' denotes the global features. 'L+G' represents local features and global features. FI2 denotes the binary sentiment classification results on the FI dataset, and FI8 denotes the classification results of eight emotions on FI (the same below).

Dataset	**WSCNet** [12]			**PDANet** [111]		
	L	G	L+G	L	G	L+G
FI2	0.894 (3)	0.894 (3)	0.896 (1)	0.807 (3)	0.876 (1)	0.878 (1)
FI8	0.671 (3)	0.675 (2)	0.679 (1)	0.606 (3)	0.696 (2)	0.694 (1)
Flickr_LDL	0.697 (3)	0.707 (2)	0.709 (1)	0.592 (3)	0.703 (1)	0.703 (1)
Twitter_LDL	0.764 (2)	0.773 (1)	0.766 (2)	0.725 (3)	0.762 (2)	0.763 (1)
Comics	0.531 (3)	0.532 (2)	0.542 (1)	0.263 (3)	0.595 (1)	0.588 (2)
GAPED	0.899 (2)	0.889 (3)	0.919 (1)	0.697 (3)	0.939 (2)	0.950 (1)
Event	0.938 (2)	0.937 (3)	0.948 (1)	0.791 (3)	0.937 (2)	0.946 (1)
Flickr	0.800 (3)	0.801 (2)	0.807 (1)	0.757 (3)	0.808 (2)	0.819 (1)
Instagram	0.804 (3)	0.816 (1)	0.804 (3)	0.672 (3)	0.811 (1)	0.807 (2)
Twitter I	0.819 (3)	0.827 (1)	0.827 (1)	0.606 (3)	0.839 (2)	0.858 (1)
Twitter II	0.824 (1)	0.824 (1)	0.815 (3)	0.815 (1)	0.815 (1)	0.815 (1)
Average rank	2.545 (3)	1.909 (2)	1.778 (1)	3.444 (3)	1.889 (2)	1.444 (1)

4.2.3 Comparison Between Local and Global Features

To fairly evaluate the effectiveness of local features and global features, we conduct the comparison experiments in Table 5 based on WSCNet and PDANet, which are the state-of-the-art methods that consider local regions. In WSCNet and PDANet, the final representation consists of both local features and global features. We conduct the comparison experiments for local, global, and combined local-global features for the two representative methods. In the last row of the table, we provide the results of average rank, which demonstrate that the global features outperform the local features in general. Besides, for most datasets, the results using both local and global features are better than that using only one type of features. It is mainly because both local and global features can determine the emotions to some extent, and some local regions may generate emotional prioritization effect rather than sole effect. Therefore, local features should be effectively integrated with global features for the more discriminative representations [115].

4.3 Quantitative Feature Comparison

In Table 6, we evaluate the performance of different features based on different classifiers. The results of six representative widely-used datasets are reported. Note that FI is regarded as a dataset that simultaneously has two sentiment categories and eight emotion categories. The hand-crafted features include PAEF [9], Sun attribute [7], and SentiBank [10], while off-the-shelf deep features are extracted from MVSO [78] and pre-trained VGGNet-16 [114]. Each type of feature is used to train three classifiers, including kNN, Naive Bayes (NB), and support vector machine (SVM). In Sun attribute, we extract the four types of features, including GIST, HOG 2×2, self-similarity, and geometric context color histogram features. In pre-trained VGGNet-16, we extract 4096-dimensional features from the last layer. Note that the features of Sun attribute and pre-trained VGGNet-16 are both reduced to 256-dimension. We report the average results of different classifiers for the same feature to fairly investigate the representation ability of each feature. From the results of the same classifiers and the average results of different classifiers, it is observed that deep features obtain the best performance, which is also demonstrated

TABLE 6
Experimental results of different features on widely-used datasets. For each feature, the average results of different classifiers are also reported.

Dataset	PAEF [9]				Sun [7]				SentiBank [10]				MVSO [78]				P-VGG [114]			
	kNN	NB	SVM	Avg	kNN	NB	SVM	Avg	kNN	NB	SVM	Avg	kNN	NB	SVM	Avg	kNN	NB	SVM	Avg
Emotion6	0.246	0.288	0.359	0.298	0.268	0.323	0.306	0.299	0.283	0.290	0.342	0.305	0.431	0.460	0.508	0.466	0.429	0.453	0.510	0.464
FI2	0.687	0.733	0.730	0.717	0.593	0.604	0.706	0.634	0.603	0.815	0.815	0.744	0.797	0.706	0.831	0.778	0.820	0.737	0.851	0.803
FI8	0.286	0.299	0.343	0.309	0.207	0.145	0.253	0.202	0.445	0.288	0.506	0.413	0.529	0.389	0.600	0.506	0.556	0.497	0.630	0.561
Flickr	0.627	0.640	0.674	0.647	0.634	0.639	0.683	0.652	0.581	0.608	0.694	0.628	0.697	0.699	0.771	0.722	0.707	0.699	0.777	0.728
Instagram	0.556	0.589	0.638	0.647	0.561	0.586	0.631	0.652	0.584	0.576	0.662	0.628	0.667	0.717	0.750	0.711	0.701	0.712	0.772	0.728
Twitter I	0.593	0.633	0.675	0.634	0.565	0.615	0.643	0.608	0.526	0.564	0.602	0.564	0.696	0.606	0.775	0.692	0.674	0.729	0.741	0.715
Twitter II	0.656	0.777	0.777	0.737	0.672	0.606	0.777	0.685	0.632	0.661	0.777	0.690	0.651	0.777	0.777	0.734	0.631	0.643	0.792	0.689

in traditional computer vision tasks, such as image classification and object detection. Besides, high-level features (*e.g*. SentiBank) perform better than middle- and low-level features (*e.g*. PAEF and Sun) in most cases. It is mainly because high-level features are more related to the emotional semantics. For example, SentiBank is constructed based on ANPs, where adjective can be better mapped into sentiment.

5 Learning Methods for Different Tasks

In this section, we review the learning methods of recent two decades on AICA, in which significant development has been obtained on different AICA tasks, including dominant emotion recognition, personalized emotion prediction, emotion distribution learning, and learning from noisy data or few labels.

5.1 Dominant Emotion Recognition

5.1.1 Traditional Methods

In the early years, researchers mainly used SVM to classify images based on various hand-crafted emotional features. Machajdik and Hanbury [8] combined the features on different levels to generate the final emotion representation. The experiments are conducted using SVMs on three small datasets using a 5-fold cross validation, and each class is separated against the others in rotation. The results are reported by true positive rate per class. The 1,200 ANPs concept detectors are trained by SVMs, resulting in SentiBank [10], which are the crucial high-level cues for sentiment prediction due to strong co-occurrence relation with sentiments. As the extensions of SentiBank [10], DeepSentiBank [103] and MVSO [78] train the detectors for 2,089 and 4,342 English ANPs, respectively, using existing deep architecture like CaffeNet, and then, the sentiment polarity can be inferred. Using text parsing technology and lexicon-based sentiment analysis tools, the adjectives can be mapped into "positive" or "negative"; likewise, the polarity of an image is derived. Even after the emergence of CNNs, SVMs also serve as an essential classifier. For instance, Ahsan et al. [83] detected event concepts through a trained CNN model and map the visual attributes into specific sentiments based on an SVM classifier. Besides, hand-crafted art features and CNN features have been combined to generate final representations [116], which are then input into SVMs for classification.

Inferring the evoked emotion from art paintings has been a significant research problem in recent years. Due to the abstract style, recognizing the emotions of art paintings becomes a very challenging task. Later, considering that an image may be represented in various feature spaces, multiple kernel learning [117] is employed to capture the different emotional patterns of abstract art. In the process, the weights of different features can be adjusted automatically, so that the learned feature combination is the most suitable one. Intuitively, the emotion is relevant to various elements of paintings like the painting technique. Therefore, non-linear matrix completion (NLMC) [76] is introduced as a transductive classifier to model the relations between different latent variables. This work well imitates the process of inferring emotion from art paintings. To tackle the scarcity of well-labeled paintings, Lu et al. [118] proposed an adaptive learning to use the labeled photographs and unlabeled paintings to identify the emotions of paintings. The differences between the two types of images are considered in the learning process.

TABLE 7
Average results of different features (PAEF [119], Sun [7], SentiBank [10], MVSO [78], pre-trained VGGNet-16 [114]) for the same classifier.

Dataset	kNN_Avg	NB_Avg	SVM_Avg
Emotion6	0.331	0.363	0.405
FI2	0.700	0.719	0.787
FI8	0.405	0.323	0.466
Flickr	0.649	0.657	0.720
Instagram	0.614	0.636	0.691
Twitter I	0.611	0.629	0.687
Twitter II	0.648	0.693	0.780

For the same classifier, we compute the average performance of different features as shown in Table 7. Generally, SVM obtains the best performance in the three classifiers, while kNN performs worse than the others except the results on FI8.

5.1.2 Learning-based Methods

Benefiting from the strong ability of CNNs to extract features, an increasing number of studies [120, 121, 12, 122] design various learning-based methods to recognize image emotions. In the early studies, a CNN is often directly used as the off-the-shelf tool without any modification. For example, Xu et al. [30] trained two classifiers following the two fully connected (FC) layers (FC7 and FC8) of an existing basic network (AlexNet), respectively. The experimental results show that the classifier after the FC7 (0.649) layer performs better than that after the FC8 (0.615). It demonstrates that the 7th layer of CNN characterizes more sentiment information of image than object detection scores in the 8th layer. To further gain insight about the influence of CNN patterns on visual sentiment analysis, Campos et al. [121] [123] gave a layer-by-layer analysis of a fine-tuned CaffeNet based on both softmax and SVM classifiers.

Personalized Network. With the further development of CNNs, more and more researchers begin working to build novel networks for better emotion recognition performance, guided by the theories of art and psychology. In [124], Wang et al. proposed a deep coupled adjective and noun neural network to recognize positive and negative sentiment from images. The architecture consisting of two parallel sub-networks (A-net and N-net) can jointly predict the adjectives and nouns of ANPs. When ANP labels are unavailable, a mutual supervision is proposed to predict the expected output

TABLE 8
Experimental results of learning-based methods on widely-used datasets. The backbone of these methods is replaced with different architectures, including AlexNet (Alex), VGGNet-16 (VGG), ResNet-50 (Res), and Inception-v3 (Inc). The average results of different backbones are reported.

Dataset	**DCNN** [30]					**RCA** [107]					**WSCNet** [12]					**PDANet** [111]				
	Alex	VGG	Res	Inc	Avg	Alex	VGG	Res	Inc	Avg	Alex	VGG	Res	Inc	Avg	Alex	VGG	Res	Inc	Avg
Emotion6	0.489	0.483	0.532	0.546	0.512	0.512	0.530	0.546	0.559	0.537	0.463	0.514	0.551	0.495	0.506	0.488	0.517	0.569	0.487	0.515
FI2	0.828	0.868	0.882	0.885	0.866	0.830	0.870	0.872	0.887	0.865	0.828	0.868	0.896	0.875	0.867	0.828	0.867	0.888	0.881	0.866
FI8	0.576	0.614	0.660	0.668	0.630	0.559	0.641	0.668	0.678	0.637	0.586	0.650	0.679	0.668	0.646	0.589	0.664	0.661	0.663	0.644
Flickr	0.737	0.770	0.760	0.784	0.763	0.786	0.813	0.800	0.801	0.800	0.776	0.783	0.811	0.786	0.789	0.779	0.794	0.780	0.793	0.787
Instagram	0.715	0.770	0.780	0.794	0.765	0.759	0.803	0.784	0.788	0.784	0.752	0.785	0.804	0.777	0.780	0.746	0.798	0.807	0.785	0.784
Twitter I	0.791	0.811	0.829	0.810	0.810	0.802	0.834	0.825	0.821	0.821	0.772	0.805	0.826	0.814	0.804	0.782	0.698	0.833	0.846	0.790
Twitter II	0.714	0.724	0.745	0.784	0.742	0.656	0.648	0.737	0.774	0.704	0.789	0.807	0.812	0.785	0.798	0.724	0.777	0.777	0.799	0.762

of each sub-network using a transition matrix that captures the relation between noun and adjective.

Multi-level Features. To fully leverage the multi-scale features of image as in [104], Zhu et al. [105] integrated the CNN and RNN architectures. Specifically, a CNN is used to extract features from different levels, and then, a bidirectional gated recurrent unit (Bi-GRU) captures the dependency among them. Finally, the two outputs from Bi-GRU are concatenated for emotion classification. In the learning process, softmax loss and contrastive loss are both used for training the model. With the contrastive loss, the features extracted from the images of the same category are enforced to be close to each other, while the features extracted from the images of different categories are enforced to be far away with each other. This study is the first to model the relations between features on different levels dynamically.

Emotional Polarities. In Mikels' eight basic emotions, there exist two polarities: *positive* and *negative*. The emotions in the same polarity are closer to each other—hence, they are highly related. This characteristic of emotion has been focused upon in several studies. Based on the triplet loss [125], Yang et al. [107] took the characteristic of polarity into account and designed a sentiment metric loss, in which the quadruplet $\{$*anchor, positive, related, negative*$\}$ is constructed for learning, where *related* denotes the sample that belongs to the same polarity with the *anchor* but different categories. By jointly optimizing softmax loss and sentiment metric loss, the architecture can be used for both classification and retrieval tasks. He and Zhang [126] designed a unified architecture consisting of two parts: a sub-network for sentiment polarity classification and a sub-network for specific emotion classification. With the assisted learning strategy, the results of the polarity can be used as important prior knowledge for more fine-grained emotion analysis. Yao et al. [113] designed emotion-pair loss by considering hierarchical structure in emotions. Based on the metric learning strategy, the features of the samples from the same polarity will be closer to each other in embedding space. It is beneficial to rank the images according to the emotional similarity with the given image.

Local Information. In complex images, some informative regions may become crucial elements to determine the evoked dominant emotion [85, 127]. Therefore, the studies by detecting regions for better recognition performance emerge rapidly. Sun et al. [128] [110] exploited an off-the-shelf objectness tool to generate proposals, and then computed the object and sentiment scores to select top regions from the masses of candidates. The selected regions are aggregated with the whole image for more discriminative representation used in emotion classification. Based on [110], Wu et al. [129] employed salient object detection model to capture informative regions, and then fed both sub-images and entire images into the network for extracting local and global features. With the same backbone (VGG-16), [129] outperforms [110] on all the commonly used datasets. It is reasonable to infer that [129] captures more discriminative information by inputting the cropped original images into network. Obviously, the above methods are time- and computing-consuming when selecting informative regions. Later, WSCNet [12] is developed to automatically generate an attention map in a single shot based on the response on feature maps, saving considerable amounts of time and computational resources. Note that each attention map is obtained by computing the weighted sum of the activation for each class. In [130], a novel fourth channel, named focal channel, is added in neural networks by taking the focal object mask of the image or the saliency map as input. By encoding the local information for sentiment representation, it is shown that negative sentiment is mainly evoked by the focal region and hardly influenced by context, whereas positive sentiment is decided by both focal region and context. A Sentiment Network with visual Attention (SentiNet-A) is proposed in [131], where the attention distributions of spatial regions are generated. The saliency map is then derived from a multi-scale fully convolutional network (FCN) to refine the attention distribution. Recently, how to capture the emotional relation between different regions is a hot topic. Zhang et al. [132] modeled the correlation between object semantics in different image regions to infer the image sentiment based on Bayesian networks. A multi-attentive pyramidal model is proposed in [133] to extract local features at various scales, and then, a self-attention mechanism is employed to mine the relations between features of different regions.

Knowledge of Other Fields. Studying emotion recognition can also utilize knowledge from other fields. Considering the correlation between aesthetics and emotion of images, Yu et al. [134] designed a novel unified aesthetics-emotion hybrid network (AEN) to simultaneously conduct image aesthetic assessment and emotion recognition. Inspired by the emotion of the generation process in brain, Zhang et al. [135] developed a multi-subnet neural network to simulate the generation of specific emotional signals and the process of signal suppression in brain neurons.

Quantitative Comparison of Representative Deep Methods. As shown in Table 8, we conduct experiments to fairly compare four representative learning-based methods, including DCNN [30], RCA [107], WSCNet [12], and PDANet [111]. We replace the original backbone with four different backbones to evaluate the effectiveness and robustness, including AlexNet [136], VGG-16 [114], ResNet-50 [18], and Inception-v3 [137]. It is observed that the robustness of different methods is different. Compared to WSCNet and PDANet, RCA is more robust when using different architectures as backbones, because the results of RCA on each dataset fluctuate less than other methods, especially on Flickr and Emotion6. In RCA, the final image representation contains

features from multiple layers, leading to richer information, of which distinctive ability does not decrease dramatically when using shallower networks. By contrast, the methods that are only based on features of the final layer are more sensitive to the depth of the used backbones. Generally, the results of recent studies, such as RCA, WSCNet, and PDANet, are better than that of DCNN, which does not contain specialized components designed for emotion classification. Under the common experimental settings, including the same input size, initialization, backbone, *etc*., the overall results of RCA, WSCNet, and PDANet do not have distinct disparity. For different datasets, the methods that obtain the best performance are different. For example, for one small-scale dataset Twitter II (only 603 images), the methods (WSCNet and PDANet) considering local informative features perform better than others.

5.2 Personalized Emotion Prediction

Yang et al. [31] first proposed to predict emotion of social images for individuals based on user interest and social influence. The user interest is modeled by considering both text and images, the emotions of which are predicted by constructing a personalized dictionary and clustering basic color features. The social influence is measured by the emotion similarity of different users towards the same microblog. The weights of user interest and social influence are obtained by mining users' historical behaviors. Later, Rui et al. extended the weighting strategy with a probabilistic graphical model [138]. Latent from the user's historical behaviors, a set of parameters in the graph model are used to estimate the importance of content and influence. However, there are some limitations of these methods. First, the extracted visual features are very simple, which cannot well reflect the visual content. Second, several important factors are not considered, such as the temporal evolution. Third, the higher-order correlations among users and images are not well modeled.

In [15, 139], Zhao et al. made several improvements to address these issues when predicting the personalized emotions (see Figure 4 (b)) of a specified user after viewing an image, associated with online social networks. Different types of factors that may influence the emotion perception are considered: the images' visual content, the social context related to the corresponding users, the emotions' temporal evolution, and the images' location information. Rolling multi-task hypergraph learning is presented to jointly combine these factors. Each hypergraph vertex is a compound triple (u, x, S), where u represents the user, x and S are the current image and the recent past images, termed as 'target image' and 'history image set', respectively. Based on the 3 vertex components, different types of hyperedges are constructed, including target image centric, history image set centric, and user centric hyperedges. Visual features (Gist, Elements, Attributes, Principles, ANP, and Expressions) in both the target image and the history image set are extracted to represent visual content. User relationship is exploited from the user component to take social context into account. Past emotion is inferred from the history image set to reveal temporal evolution. Location is embedded in both the target image and the history image set. Semi-supervised learning is then conducted on the multi-task hypergraphs to classify personalized emotions for multiple users simultaneously.

5.3 Emotion Distribution Learning

Label distribution learning (LDL) [140] is used to model the relative importance of each category for an image. The sum of the probability on each label in discrete space is 1. It is usually used to solve the ambiguity of emotion in discrete label space.

Peng et al. [14] constructed the Emotion6 database that is annotated with probability distribution. SVR, CNN, and CNN regression (CNNR) are employed as the emotion regression model. Particularly, one SVR and CNNR are trained for each category, while one CNN is trained for all categories by changing the number of output neurons to the number of emotion categories. Zhao et al. [32, 141] modeled the emotion distribution prediction as a shared sparse learning (SSL) problem. The input is the combination of different types of features, and the objective is iteratively optimized by reweighted least squares. Later in [61], the weighted multi-modal shared sparse learning (WMMSSL) is proposed, in which the weight of different features can be learned automatically. Based on the conditional probability neutral network (CPNN), in [140], BCPNN [16] is proposed by representing the image label with binary encoding rather than the general signless integers. Besides, ACPNN is developed based on BCPNN by adding noises to the ground truth label. With this strategy, the emotion distribution is augmented, which benefits training more robust models. Zhao et al. [142] proposed a Weighted Multi-Modal Conditional Probability Neural Network (WMMCPNN) to explore the optimal combination coefficients of different types of features. In [143], Yang et al. designed a unified framework to optimize the Kullback-Leibler (KL) loss and softmax loss, simultaneously. Besides, considering the lack of manually annotated emotion distribution in some datasets, a scheme that converts the single emotion into a probability distribution is proposed in this study.

Considering the co-occurrence and mutual exclusion of some emotions, it is important to model the relation of different emotional labels when predicting the probability labels. He and Jin [144] used Graph Convolutional Networks (GCN) to model the label relationship for label distribution prediction. In detail, the GloVe-300 word embeddings of emotions are input into GCN as nodes, and the relation of different labels is computed using the probability of co-occurrence of two emotions. Liu et al. [145] integrated low-rank and inverse-covariance regularization terms into one framework for emotion distribution learning. The low-rank regularization term is used to learn low-rank structured embedding features, while the inverse-covariance regularization term can ensure the structured sparsity of regression coefficients. To fully employ the polarity and character of the intensity in emotions, structured and sparse annotations are leveraged to learn an emotion label distribution in [146].

In the continuous emotion space, Zhao et al. [17] modeled the continuous distribution with a Gaussian mixture model, in which the parameters can be estimated by the expectation-maximization algorithm. Shared Sparse Regression (SSR) is introduced as the learning model by assuming that the test feature and test parameters can be linearly represented by the training features and training parameters but with shared coefficients. To explore the task relatedness, multi-task SSR is further proposed to simultaneously predict the parameters of different test images by using proper shared information across tasks.

5.4 Learning from Noisy Data or Few Labels

Few-shot or Zero-shot Learning. As stated in Section 1.1, few/zero shot learning and unsupervised/weakly supervised learning are two possible solutions to address the label absence challenge. Few/zero shot learning refers to a specific machine learning,

where a model is learned based on very few or even no labeled examples [38]. Although humans can learn through only a small number of samples, it is difficult for machine learning to do so. Conventional methods usually construct a shared space for both seen and unseen classes. For seen classes, the space is learned based on the correspondence between the seen images and their labels. Relying on the side information (*e.g.* attributes), the unseen classes are first related to the seen classes and then mapped to the common space based on the cross-modality similarity between visual features and class semantic representations. The existence of the affective gap makes it difficult to compute this similarity.

Wang et al. [147] proposed an emotion navigation framework using auxiliary noisy data and employed the few-shot precise samples as the prototype center to guide noisy data clustering. Zhan et al. [38] proposed an affective structural embedding framework, which constructs an intermediate embedding space using ANP features for zero-shot emotion recognition. In addition, an affective adversarial constraint is introduced to select the embedding space that simultaneously preserves the affective structural information and retains the discriminative capacity.

Unsupervised/Weakly-supervised Learning. Unsupervised learning aims to find previously unknown patterns in a dataset without pre-existing labels. Two main methods are cluster analysis and principal component analysis. How to automatically determine the number of clusters is a key challenge in clustering. Differently, Wang et al. [148] exploited the relations among visual content and relevant textual information for unsupervised sentiment analysis of social images. This method relies on the accompanying text of social images. On the one hand, the text may be incomplete and noisy. On the other hand, there may be no available text. In such cases, how unsupervised analysis can be done in such cases appears worth studying.

For social images, a more practical scenario is that they are weakly and noisily labeled [102, 149, 150, 19]. Considering that the images of the VSO dataset are weakly labeled with noises, Wang et al. [102] estimated the noise matrix to reweight the softmax loss that can compensate the degeneration of classification performance resulting from the noisy labels. Retrained by reweighting the loss, the learned model is more discriminative for emotional images. Wu et al. proposed to refine the weakly labeled dataset based on the sentiments of ANPs and provided tags [149]. The images are removed if the sentiments of ANPs and tags are contradicting each other and if the numbers of positive tags and negative tags are equal. The remaining images are relabeled with the dominant sentiment of the tags. Using the refined dataset, a better performance can be obtained. Chen et al. employed a probabilistic graphical model to filter out the label noise [150]. Wei et al. [19] proposed to train a joint text and visual embedding to reduce noise in the webly annotated tags by text-based distillation. Designing an effective strategy that can better refine the dataset or filter the label noise is expected to improve the performance.

Domain Adaptation/Generalization. Domain adaptation studies how to transfer the models trained on a labeled source domain to another sparsely labeled or unlabeled target domain. One direct solution is to translate the source images to an intermediate domain that is indistinguishable from the target images while preserving the source labels using GANs [151, 152, 153]. Some existing unsupervised domain adaptation methods on AICA are based on this intuition. Zhao et al. [35] studied the domain adaptation problem in emotion distribution learning. They develop an adversarial model, termed EmotionGAN, by alternately optimizing the GAN loss, semantic consistency loss, and regression loss. The semantic consistency loss guarantees that the translated intermediate images preserve the source labels. Since traditional GANs are unstable and prone to failure [152], the cycle-consistent GAN (CycleGAN) was designed. Based on CycleGAN, Zhao et al. enforced semantic consistency when adapting the dominant emotions without requiring aligned image pairs [36, 37]. He and Ding proposed a discrepancy-based domain adaptation method [154]. Both marginal and joint domain distribution discrepancies at fully-connected layers are reduced by minimizing the joint maximum mean discrepancy. Without generating an intermediate domain, this method aims to extract more transferable features.

All the above methods focus on a single-source scenario. However, in practice, the labeled data may be collected from multiple sources with different distributions. Simply combining the multiple sources into one source and performing single-source domain adaptation may lead to suboptimal solutions. In [155], Lin et al. studied multi-source domain adaptation for binary sentiment classification of images. Specifically, a multi-source sentiment generative adversarial network (MSGAN) is designed to find a unified sentiment latent space where the source images and target images share a similar distribution. MSGAN includes three pipelines: image reconstruction, image translation, and cycle reconstruction. The results demonstrate that exploring the complementarity of multiple sources can improve the adaptation performance to a large margin as compared to best single-source adaptation methods.

Differently, Panda et al. [22] studied the domain generalization problem of AICA to overcome dataset bias. A weakly-labeled large-scale emotion dataset is constructed by collecting images from a stock website to cover a wide variety of emotion concepts. A simple yet effective curriculum guided training strategy is proposed to learn discriminative emotion features, which demonstrate better generalization ability than the existing datasets.

6 AICA Based Applications

With the booming development of AICA, the related application has been or will be on the agenda in different directions, including opinion mining, business intelligence, psychological health, and entertainment assistant, to name but a few in more detail.

6.1 Opinion Mining

Nowadays, an increasing number of people use images to express their viewpoints or attitudes towards some events. Based on the analysis of these shared images, we can infer the emotions of the different users, including uploaders and commentators. Furthermore, we can conjecture their attitudes towards the specific events or products. In [15], the different types of factors, including visual content, social context, temporal evolution and location influence, are modeled using a hypergraph model to iteratively optimize the personal social image emotion prediction. Furthermore, various virtual groups are formed according to the interests or backgrounds of users. Analyzing group-based emotions will contribute to predicting the tendency of the society. Based on the above technologies, we can imagine that the understanding of social image emotion can be used in public opinion analysis and related applications.

In special domains like product comments, the experience of users has been investigated and evaluated based on emotions from

uploaded images. In [156], Truong and Lauw conducted visual sentiment analysis for better understanding of review images about different products, services, and venues. In the process, both user and items factors are taken into account. Ye et al. [157] jointly employed a visual and textual classification to analyze the sentiment of the product reviews. Besides, a dataset named Product Reviewes-150K (PR-150K) is constructed. In [158], Hassan et al. analyzed the sentiments evoked from disaster-related images by taking into account people's opinions, attitudes, feelings, and emotions. The study sets a baseline for the future research in disaster-related images sentiment analysis. Therefore, it is significant to mine the positive or negative aspects for opinion of the users by analyzing the emotions of related images.

6.2 Psychological Health

With the popularity of the social media, people share their mood on the Internet rather than with their real friends. For a user that shares negative information continuously, it is necessary to further track her/his mental status to prevent the occurance of psychological illness and even suicide. Guntuku et al. [159] revealed how a twitter profile and post images reflect depression and anxiety. In [160], an automatic stress detection model is proposed for social web users by analyzing the emotional content of multimodal microblog data. Based on the model, we can further design subsequent decompressing services for users, including playing some smoothing music, playing some funny videos, and providing some forms of exercises, *etc*.

In the field of psychology, affective images are employed to conduct some studies. For instance, IAPS [28] is a database of images constructed to provide a standardized set to evoke a target emotion in people for studying psychological status. Each image has listed the average ratings of the elicited emotions, and these ratings can be used for various research directions in psychological theories. In [161], a new image system named Tsinghua psychological image system (ThuPIS) is built based on the Minnesota multiphasic personality inventory (MMPI) [162], which is a famous personality diagnosis tool for clinical mental health. The system can be applied to support the new psychological test for monitoring the mental health of humans.

6.3 Business Intelligence

Images play an essential role in conveying the business information, so selecting the images with proper emotions can benefit the development of a business. For example, most advertisements are presented using visual content to evoke strong emotional stimulus in viewers. Consumer research [163] has proven that emotions can affect the process of decision making. A well-designed advertisement can attract people's attention and evoke positive emotions in viewers, so that a desire of purchasing will be produced when viewing an accordingly tailored advertisement. Holbrook and O'Shaughnessy [164] investigated the role of emotion in advertising. Specifically, they distinguish emotion from other types of consumer responses, and study the emotion generating process from the emotional content in the advertisements. Besides, suggestions are put forward for the design of advertisement considering emotional elements in the future. Poels and Dewitte [165] reviewed and updated the measuring methods for emotions in advertising, and further discuss their applicability. Finally, the influence of emotions on the effectiveness of advertising is investigated.

In the filed of tourism, emotion is an important element that cannot be ignored for evaluating the overall experience of a trip [166]. By analyzing the uploaded travel photos in the social networks, the relations among motivation, image dimension, and emotional qualities of places are explored in [167]. The paper reveals that the natural resources, including "flora and fauna", "countryside", "beaches", *etc*., are always associated with the feelings of "arousing" and "pleasant" for the specific destination. Besides, the travel photos taken in a long shot, at eye-level, with stark density level can elicit happiness feelings. These findings can guide to exhibit the more attractive travel photos on some specific platforms, so as to initiate successful marketing efforts and promote the booming of tourism. In [168], a survey is conducted on the emotional experience of tourists by distributing self-administered questionnaires. The emotion feedback (*arousal* and *pleasure*) for each destination is plotted on a corresponding two-dimensional grid. Hosany and Prayag [169] empirically investigated the patterns of emotional response from tourists and discussed the relationship between these emotional patterns and the consumption satisfaction. Five different emotional response patterns (*delighted*, *unemotional*, *negative*, *mixed*, and *passionate*) are derived by cluster analysis based on a four-dimensional emotional space defined by love, surprise, joy, and unpleasantness. It is reported that these five patterns are different in the satisfaction level and the intention of recommendation. In the future, based on a more fine-grained emotion analysis, we can construct personalized destination recommendation systems for users who intend to have different travel experiences automatically.

6.4 Entertainment Assistant

Nowadays, the standard of entertainment has treated the emotion as a crucial element that can decide the entertainment experience [170]. Simultaneously, the emotion can also be used to evaluate the experience of entertainment. For instance, emotions can be regarded as the medium that links different modalities of data, such as image and music. In [171], an emotion-driven cross-media retrieval system is designed based on differential and evolutionary-support vector machine (DE-SVM). The system can achieve the retrieval between Chinese folk music and Chinese folk image based on their involved emotions. Chen et al. [172] and Zhao et al. [173] designed a system that computes the emotional similarity between music and images. With this system, users can generate the mood-aware music slide shows from their personal album photos.

Emotions in comics play a crucial role in attracting people. The reason why Indonesian readers widely accept Japanese comics has been investigated in [174]. The report indicates that the comics are not only an entertainment for them, but also a significant life experience, in which emotion is an important element. Therefore, we should take into account the evoked emotions when measuring comics. In [82], a large-scale comics dataset is constructed, in which the images are labeled with the emotions defined in Mikel's wheel. With more attention paid to AICA in the entertainment domain, we can establish a conversation with chatbots based on various types of images rather than only based on text.

7 Future Directions

Although remarkable progress has been made on affective image content analysis (AICA), there are still several open issues and directions that are worth investigating by jointly considering the

efforts from different disciplines, such as psychology, cognitive science, multimedia, and machine learning.

7.1 Image Content and Context Understanding

As emotions may be directly evoked by the image content in viewers, accurately analyzing what is contained in an image can significantly improve the performance of AICA. As stated in Section 4, there are different kinds of emotion features. Although the deep ones generally outperform the hand-crafted ones, it is unclear whether combining hand-crafted ones with deep ones can boost the performance. If yes, how to effectively fuse them? Further, the correlation between deep features and specific emotions is unclear, while the hand-crafted features—especially mid-level and high-level ones—are more understandable. Using hand-crafted features to guide the generation of interpretable deep ones is an interesting topic. Sometimes, we even need subtle analysis of image contents. For example, we may feel "happy" about beautiful flowers; but, if the flowers are placed in a funeral, we possibly feel "sad". If an image is about the laugh of a lovely child, it is more likely that we feel "amused"; but if it is about the laugh of a known evil ruler or criminal, we may feel "angry".Constructing a large-scale repository and collecting sufficient corresponding images can help to solve this problem. As shown in Fig. 3, the context of an image is also very important. Multi-modal emotion recognition would make more sense, such as textual-visual data [175, 150] and audio-visual data [176]. One key challenge is how to fuse the data of different modalities.

7.2 Viewer Contextual and Prior Knowledge Modeling

The contextual information of a viewer watching images can significantly influence the emotion perception. The same viewer can experience different emotions for the same image depending on the context, such as climate, time, and social context [15]. Incorporating these important contextual factors can be expected to boost the performance. Using probabilistic graph or hypergraph models to represent the complex correlations of different factors is demonstrated to be feasible [79, 177, 15]. We may further try to model these factors by more recent graph convolutional networks [178] and hypergraph neural networks [179].

The prior knowledge of viewers, such as gender and personality, may also influence the emotion perception. For example, an optimistic viewer and a pessimistic viewer may have totally different emotions about the same image. Wu et al. investigate the influence of user demographics, including gender, marital status, and occupation, as related to the emotion perception of social images [180, 181]. Besides the visual content, temporal correlation, and social correlation, user demographics are also incorporated as factor functions in a factor graph model. The results show that user demographics can indeed improve the overall emotion classification performance. However, the collected prior knowledge on social networks may be inaccurate. How to automatically filter the noisy ones has not been investigated.

7.3 Learning from Noisy Data or Few Labels

Few-shot or Zero-shot Learning. There are some limitations of current few-shot/zero-shot learning methods for AICA [147, 38]. First, not all seen images are helpful in generating the embedding space. How to automatically select the representative images to generate better embedding space is unclear. Second, the embedding process may result in information loss, and the cross-modality similarity cannot make full use of the data distribution. We may consider synthesizing reliable samples for the unseen classes based on the estimated distribution. With the success of Generative Adversarial Networks (GANs) [151], such an idea would bear genuine potential.

Domain Adaptation/Generalization. In practice, we might have a few labeled target images. In such cases, the domain adaptation task becomes a semi-supervised scenario. One interesting problem is how many labeled target images are required at least to achieve or even outperform the results fully trained on the target domain. Besides the semi-supervised domain adaptation, some other challenging problems include heterogeneous domain adaptation where the label space is different between the source and target domains, open set domain adaptation where both source and target domains contain images that do not belong to the classes of interest, and category shift domain adaptation where the categories from different sources might be different.

While the target images (although without labels) are available in domain adaptation, *i.e*. the target images are accessible during the training process, domain generalization learns a model without accessing any target image [182]. To enrich the generalization ability, one possible solution is to randomize the labeled source images to a sufficient number of domains in the training stage using domain randomization [183], and then, the target domain belongs to the randomized domains to a large extent and thus, the models trained on the randomized domain can well adapt to the target domain.

7.4 Group Emotion Clustering

Simply recognizing the dominant emotion for an image is too generic, while predicting personalized emotion for each user is too specific. Since some groups or cliques of users, who share similar tastes or interests and have similar background, are more likely to respond similarly to the same image, it would make more sense to predict emotions for these groups or cliques. Analyzing the user profiles provided by each individual to classify users into different types of groups based on gender, backgrounds, tastes, interests, and so on may provide a feasible solution.

Current research on group emotion mainly focuses on recognizing the emotions of the groups of people contained in an image attending a wide variety of social events [184, 185]. Affective image analysis for groups of people, *i.e*. recognizing the induced emotions of the groups, has not been explored yet. Group emotion recognition plays an important role in recommendation. For example, for the people in the same group, if one is interested in a specific product, the others are more likely to accept it.

7.5 Viewer-Image Interaction

Besides the direct analysis of image content, we may also record and analyze the viewers' audiovisual or physiological responses when watching the image (such as facial expressions, or electroencephalogram signals), which is often called implicit emotional tagging.Current methods mainly focus on videos [186, 187, 188, 189] for its relative emotional consistency temporally. Exploring viewers' responses for implicit emotion analysis of images is still a largely open topic of research. Jointly modeling both image content and viewers' responses may better bridge the affective gap and result in superior performances. In practice, some data

may be missing or corrupted. For example, some physiological signals are not successfully captured. In such cases, how to deal with missing data should be considered.

As explained in Section 1, the physiological responses are either difficult to capture or easily suppressed. In real-world applications, even if there are no physiological responses, jointly exploring the privileged modality during training might also lead to better performance than using the image modality only.

7.6 Novel and Real-world AICA-based Applications

With the availability of large-scale datasets and improvements in machine learning, especially in deep learning, the AICA performance will be significantly boosted. Therefore, we foresee the coming of an emotional intelligence era with more AICA-based real-world applications. For example, in online fashion recommendation, intelligent costumer services, such as customer-image interaction, can provide better experience to customers. In advertisement, generating or curating images that can evoke expected emotions strongly can attract more attention. One preliminary image adjustment system is implemented in [190]. Given an input image and an affective word, the system can adjust image color to meet the desired emotion. Only the color information is changed, which may be insufficient in applications. Peng et al. instead proposed to modify the evoked emotion distribution of the given source image towards that of the target image by changing color and texture related features [14]. We believe that GAN-based adversarial models are possibly more suitable to generate affective images. In art theory, we can understand how artists express emotions through their artworks. The principles can guide the affective image generation. The generated synthetic images can in turn improve the AICA results through domain adaptation. In education, the images with enriched emotions can help children to better learn and understand. Certainly, many more exciting applications will be coming up soon.

7.7 Efficient AICA Learning

There are three factors that attribute to the success of deep learning: increased computing capacity, deep complex models, and sufficient labeled data. However, these factors may be unavailable for edge devices such as mobile phones which are widely used in our daily life but have limited power, memory, and computing capacity. Therefore, designing specialized and efficient "green" deep learning models is required. Efficient model design has been actively studied in computer vision. Some efficient representation methods include auto channel pruning, student-teacher network approaches, neural network and hardware accelerator co-design, auto mixed-precision quantization, optimal neural architecture search, *etc*.

To the best of our knowledge, the efficiency problem has not been well studied in AICA. Extending existing methods in computer vision to the AICA task by incorporating its speciality (*e.g*. emotion hierarchy) is a simple but effective solution. It would make more sense if the on-device training models can learn online with incremental data.

7.8 Benchmark Dataset Construction

The datasets adopted in existing AICA studies are mainly well-labeled small-scale ones (*e.g*. IAPSa [29]) or large-scale ones with labels obtained by a keyword searching strategy (*e.g*. IESN [15]). While there are not enough training samples in the former ones, the label quality of the automatic annotations cannot be guaranteed in the latter case. Creating a large-scale and high-quality dataset, like the ImageNet in computer vision, can significantly advance the development of AICA. One possible solution is to exploit online systems and crowd-sourcing platforms to invite/attract large numbers of viewers with a representative spread of backgrounds to annotate their personalized emotion perceptions of images together with the contextual information on their emotional responses. Personalized emotion annotation would better accord with the subjectiveness of emotions. Further, from the personalized emotions, we can obtain both the dominant emotion and emotion distribution. Collecting the social media users' interaction with images, *e.g*., likes, comments, together with their spontaneous responses, *e.g*., facial expression, where possible, can provide more information to enrich affective datasets. To facilitate the applicability of AICA in practice with different emotion requirements, employing a hierarchical model (*e.g*. Parrott [64]) with emotion intensity is a good choice.

8 Conclusion

This article attempted to provide a comprehensive survey of recent developments on affective image content analysis (AICA) over the last two decades. Obviously, it cannot cover all the literature on AICA, and we focused on a representative subset of the latest methods. We summarized and compared the widely employed emotion representation models, available datasets, and the state-of-the-art methods on emotion feature extraction, learning methods, and AICA-based applications. Finally, we discussed some open issues and potential research directions in AICA. Although deep learning-based AICA methods have achieved remarkable progress recently, an effective, efficient, and robust AICA algorithm that can achieve satisfying performance under unconstrained conditions is yet to be designed. With the rapid development of deep understanding of emotion evocation in brain science, accurate emotion measurement in psychology, and novel deep learning network architectures in machine learning, we believe that AICA will continue to be an active and promising research topic for a long time.

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

References

[1] M. Minsky, *The Society of mind*. Simon and Schuster, 1986.
[2] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *CSUR*, vol. 47, no. 3, p. 43, 2015.
[3] S. Zhao, Y. Ma, Y. Gu, J. Yang, T. Xing, P. Xu, R. Hu, H. Chai, and K. Keutzer, "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *AAAI*, 2020, pp. 303–311.
[4] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *AAAI*, 2016, pp. 308–314.
[5] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv," *IEEE SPM*, vol. 23, no. 2, pp. 90–100, 2006.
[6] V. Yanulevskaya, J. C. van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek, "Emotional valence categorization using holistic image features," in *ICIP*, 2008, pp. 101–104.
[7] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *CVPR*, 2012, pp. 2751–2758.

[8] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *ACM MM*, 2010, pp. 83–92.

[9] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *ACM MM*, 2014, pp. 47–56.

[10] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *ACM MM*, 2013, pp. 223–232.

[11] M. Chen, L. Zhang, and J. P. Allebach, "Learning deep features for image emotion classification," in *ICIP*, 2015, pp. 4491–4495.

[12] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang, "Weakly supervised coupled networks for visual sentiment analysis," in *CVPR*, 2018, pp. 7584–7592.

[13] R. Kosti, J. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using emotic dataset," *IEEE TPAMI*, vol. 42, no. 11, pp. 2755–2766, 2020.

[14] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *CVPR*, 2015, pp. 860–868.

[15] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T.-S. Chua, "Predicting personalized emotion perceptions of social images," in *ACM MM*, 2016, pp. 1385–1394.

[16] J. Yang, M. Sun, and X. Sun, "Learning visual sentiment distributions via augmented conditional probability neural network," in *AAAI*, 2017, pp. 224–230.

[17] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, "Continuous probability distribution prediction of image emotions via multi-task shared sparse regression," *IEEE TMM*, vol. 19, no. 3, pp. 632–645, 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[19] Z. Wei, J. Zhang, Z. Lin, J.-Y. Lee, N. Balasubramanian, M. Hoai, and D. Samaras, "Learning visual emotion representations from web data," in *CVPR*, 2020, pp. 13 106–13 115.

[20] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR*, 2011, pp. 1521–1528.

[21] S. Zhao, X. Yue, S. Zhang, B. Li, H. Zhao, B. Wu, R. Krishna, J. E. Gonzalez, A. L. Sangiovanni-Vincentelli, S. A. Seshia *et al.*, "A review of single-source deep unsupervised visual domain adaptation," *IEEE TNNLS*, 2020.

[22] R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, and A. K. Roy-Chowdhury, "Contemplating visual emotions: Understanding and overcoming dataset bias," in *ECCV*, 2018, pp. 579–595.

[23] J. D. Williamson, "Speech analyzer for analyzing frequency perturbations in a speech pattern to determine the emotional state of a person," Feb. 27 1979, uS Patent 4,142,067.

[24] J. E. Cahn, "The generation of affect in synthesized speech," *Journal of the American Voice I/O Society*, vol. 8, no. 1, pp. 1–1, 1990.

[25] H. Kobayashi and F. Hara, "Recognition of six basic facial expression and their strength by neural network," in *ROMAN*, 1992, pp. 381–386.

[26] P. Salovey and J. D. Mayer, "Emotional intelligence," *Imagination, Cognition and Personality*, vol. 9, no. 3, pp. 185–211, 1990.

[27] R. W. Picard, *Affective computing*. MIT press, 1997.

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

[29] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *BRM*, vol. 37, no. 4, pp. 626–630, 2005.

[30] C. Xu, S. Cetintas, K. Lee, and L. Li, "Visual sentiment prediction with deep convolutional neural networks," *arXiv:1411.5731*, 2014.

[31] Y. Yang, P. Cui, W. Zhu, and S. Yang, "User interest and social influence based emotion prediction for individuals," in *ACM MM*, 2013, pp. 785–788.

[32] S. Zhao, H. Yao, X. Jiang, and X. Sun, "Predicting discrete probability distribution of image emotions," in *ICIP*, 2015, pp. 2459–2463.

[33] S. Zhao, H. Yao, and X. Jiang, "Predicting continuous probability distribution of image emotions in valence-arousal space," in *ACM MM*, 2015, pp. 879–882.

[34] A. Liu, Y. Shi, P. Jing, J. Liu, and Y. Su, "Low-rank regularized multi-view inverse-covariance estimation for visual sentiment distribution prediction," *JVCIR*, vol. 57, pp. 243–252, 2018.

[35] S. Zhao, X. Zhao, G. Ding, and K. Keutzer, "Emotiongan: unsupervised domain adaptation for learning discrete probability distributions of image emotions," in *ACM MM*, 2018, pp. 1319–1327.

[36] S. Zhao, C. Lin, P. Xu, S. Zhao, Y. Guo, R. Krishna, G. Ding, and K. Keutzer, "Cycleemotiongan: Emotional semantic consistency preserved cyclegan for adapting image emotions," in *AAAI*, 2019, pp. 2620–2627.

[37] S. Zhao, X. Chen, X. Yue, C. Lin, P. Xu, R. Krishna, J. Yang, G. Ding, A. L. Sangiovanni-Vincentelli, and K. Keutzer, "Emotional semantics-preserved and feature-aligned cyclegan for visual emotion adaptation," *IEEE TCYB*, 2021.

[38] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang, "Zero-shot emotion recognition via affective structural embedding," in *ICCV*, 2019, pp. 1151–1160.

[39] A. Giachanou and F. Crestani, "Like it or not: A survey of twitter sentiment analysis methods," *CSUR*, vol. 49, no. 2, p. 28, 2016.

[40] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[41] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *PR*, vol. 44, no. 3, pp. 572–587, 2011.

[42] B. W. Schuller, "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends," *CACM*, vol. 61, no. 5, pp. 90–99, 2018.

[43] F. Metze, A. Batliner, F. Eyben, T. Polzehl, B. Schuller, and S. Steidl, "Emotion recognition using imperfect speech recognition," in *INTERSPEECH*, 2010, pp. 478–481.

[44] B. Schuller, C. Hage, D. Schuller, and G. Rigoll, "'mister dj, cheer me up!': Musical and textual features for automatic mood classification," *JNMR*, vol. 39, no. 1, pp. 13–34, 2010.

[45] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM TIST*, vol. 3, no. 3, p. 40, 2012.

[46] B. Schuller, S. Hantke, F. Weninger, W. Han, Z. Zhang, and S. Narayanan, "Automatic recognition of emotion evoked by general sound events," in *ICASSP*, 2012, pp. 341–344.

[47] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE TCYB*, vol. 36, no. 2, pp. 433–449, 2006.

[48] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE TPAMI*, vol. 37, no. 6, pp. 1113–1133, 2014.

[49] S. Li and W. Deng, "Deep facial expression recognition: A survey," *arXiv:1804.08348*, 2018.

[50] T. Hassan, D. Seuß, J. Wollenberg, K. Weitz, M. Kunz, S. Lautenbacher, J.-U. Garbas, and U. Schmid, "Automatic detection of pain from facial expressions: a survey," *IEEE TPAMI*, 2019.

[51] S. Wang and Q. Ji, "Video affective content analysis: a survey of state-of-the-art methods," *IEEE TAFFC*, vol. 6, no. 4, pp. 410–430, 2015.

[52] S. Zhao, Y. Ma, Y. Gu, J. Yang, T. Xing, P. Xu, R. Hu, H. Chai, and K. Keutzer, "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *AAAI*, 2020, pp. 303–311.

[53] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using eeg signals: A survey," *IEEE TAFFC*, 2017.

[54] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller, "End-to-end learning for dimensional emotion recognition from physiological signals," in *ICME*, 2017, pp. 985–990.

[55] S. Zhao, A. Gholaminejad, G. Ding, Y. Gao, J. Han, and K. Keutzer, "Personalized emotion recognition by personality-aware high-order learning of physiological signals," *ACM TOMM*, vol. 15, no. 1s, p. 14, 2019.

[56] B. Schuller, M. Lang, and G. Rigoll, "Multimodal emotion recognition in audiovisual communication," in *ICME*, vol. 1, 2002, pp. 745–748.

[57] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *IVC*, vol. 65, pp. 3–14, 2017.

[58] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *INF*, vol. 37, pp. 98–125, 2017.

[59] S. Zhao, S. Wang, M. Soleymani, D. Joshi, and Q. Ji, "Affective computing for large-scale heterogeneous multimedia data: A survey," *ACM TOMM*, vol. 15, no. 3s, p. 93, 2019.

[60] S. Zhao, G. Ding, Q. Huang, T.-S. Chua, B. W. Schuller, and K. Keutzer, "Affective image content analysis: A comprehensive survey." in *IJCAI*, 2018, pp. 5534–5541.

[61] S. Zhao, G. Ding, Y. Gao, and J. Han, "Approximating discrete probability distribution of image emotions by multi-modal features fusion," in *IJCAI*, 2017, pp. 4669–4675.

[62] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[63] R. Plutchik, *Emotion: A psychoevolutionary synthesis*. Harpercollins College Division, 1980.

[64] W. G. Parrott, *Emotions in social psychology: Essential readings*. Psychology Press, 2001.

[65] H. Schlosberg, "Three dimensions of emotion," *Psychological Review*, vol. 61, no. 2, p. 81, 1954.

[66] J. Lee and E. Park, "Fuzzy similarity-based emotional classification of color images," *IEEE TMM*, vol. 13, no. 5, pp. 1031–1039, 2011.

[67] M. D. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE TAFFC*, vol. 5, no. 2, pp. 101–111, 2014.

[68] K. R. Scherer, A. Schorr, and T. Johnstone, *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001.

[69] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 1988.

[70] S. Zhao, G. Ding, J. Han, and Y. Gao, "Personality-aware personalized emotion recognition from physiological signals," in *IJCAI*, 2018, pp. 1660–1667.

[71] H. Gunes and B. Schüller, "16 automatic analysis of social emotions," *SSP*, p. 213, 2017.

[72] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *IVC*, vol. 31, no. 2, pp. 120–136, 2013.

[73] K. Sun, J. Yu, Y. Huang, and X. Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," in *ICME*, 2009, pp. 566–569.

[74] S. M. Alarcão and M. J. Fonseca, "Identifying emotions in images from valence and arousal ratings," *MTA*, vol. 77, no. 13, pp. 17 413–17 435, 2018.

[75] E. S. Dan-Glauser and K. R. Scherer, "The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance," *BRM*, vol. 43, no. 2, pp. 468–477, 2011.

[76] X. Alameda-Pineda, E. Ricci, Y. Yan, and N. Sebe, "Recognizing emotions from abstract paintings using non-linear matrix completion," in *CVPR*, 2016, pp. 5240–5248.

[77] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks." in *AAAI*, 2015, pp. 381–388.

[78] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang, "Visual affect around the world: A large-scale multilingual visual sentiment ontology," in *ACM*

MM, 2015, pp. 159–168.
[79] Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, C. Xing, and J. Tang, "How do your friends on social media disclose your emotions?" in *AAAI*, 2014, pp. 306–312.
[80] M. Katsurai and S. Satoh, "Image sentiment analysis using latent correlations among visual, textual, and sentiment views," in *ICASSP*, 2016, pp. 2837–2841.
[81] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, and M. Tesconi, "Cross-media learning for image sentiment analysis in the wild," in *ICCVW*, 2017, pp. 308–317.
[82] D. She, M. Sun, and J. Yang, "Learning discriminative sentiment representation from strongly-and weakly supervised cnns," *ACM TOMM*, vol. 15, no. 3s, pp. 1–19, 2019.
[83] U. Ahsan, M. De Choudhury, and I. Essa, "Towards using visual attributes to infer image sentiment of social events," in *IJCNN*, 2017, pp. 1372–1379.
[84] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," in *CVPR*, 2017, pp. 1667–1675.
[85] S. Fan, Z. Shen, M. Jiang, B. L. Koenig, J. Xu, M. S. Kankanhalli, and Q. Zhao, "Emotional attention: A study of image sentiment and visual attention," in *CVPR*, 2018, pp. 7521–7531.
[86] P. Balouchian, M. Safaei, and H. Foroosh, "Lucfer: A large-scale context-sensitive image dataset for deep learning of visual emotions," in *WACV*, 2019, pp. 1645–1654.
[87] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *BRM*, vol. 45, no. 4, pp. 1191–1207, 2013.
[88] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.
[89] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *IJCV*, vol. 127, no. 3, pp. 302–321, 2019.
[90] W.-n. Wang, Y.-l. Yu, and J.-c. Zhang, "Image emotional classification: static vs. dynamic," in *SMC*, vol. 7, 2004, pp. 6407–6411.
[91] W.-n. Wang, Y.-l. Yu, and S.-m. Jiang, "Image retrieval by emotional semantics: A study of emotional space and feature extraction," in *SMC*, vol. 4, 2006, pp. 3534–3539.
[92] H. Zhang, E. Augilius, T. Honkela, J. Laaksonen, H. Gamper, and H. Alene, "Analyzing emotional semantics of abstract art using low-level image features," in *IDA*, 2011, pp. 413–423.
[93] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang, "On shape and the computability of emotions," in *ACM MM*, 2012, pp. 229–238.
[94] A. Sartori, D. Culibrk, Y. Yan, and N. Sebe, "Who's afraid of itten: Using the art theory of color combination to analyze emotions in abstract paintings," in *ACM MM*, 2015, pp. 311–320.
[95] J. Yuan, S. Mcdonough, Q. You, and J. Luo, "Sentribute: image sentiment analysis from a mid-level perspective," in *WISDOM*, 2013, p. 10.
[96] X. Lu, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang, "An investigation into three visual characteristics of complex scenes that evoke human emotion," in *ACII*, 2017, pp. 440–447.
[97] X. Wang, J. Jia, J. Yin, and L. Cai, "Interpretable aesthetic features for affective image classification," in *ICIP*, 2013, pp. 3230–3234.
[98] T. Rao, M. Xu, H. Liu, J. Wang, and I. Burnett, "Multi-scale blocks based image emotion classification using multiple instance learning," in *ICIP*, 2016, pp. 634–638.
[99] P. Yang, Q. Liu, and D. N. Metaxas, "Exploring facial expressions with compositional features," in *CVPR*, 2010, pp. 2638–2644.
[100] A. R. Ali, U. Shahid, M. Ali, and J. Ho, "High-level concepts for affective understanding of images," in *WACV*, 2017, pp. 679–687.
[101] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE TPAMI*, vol. 38, no. 3, pp. 447–461, 2015.
[102] L. Wang, X. Xu, K. Guo, and B. Cai, "Visual sentiment analysis with noisy labels by reweighting loss," in *SMC*, 2018, pp. 1873–1878.
[103] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," *arXiv:1410.8586*, 2014.
[104] T. Rao, M. Xu, and D. Xu, "Learning multi-level deep representations for image emotion classification," *NPL*, vol. 51, no. 3, pp. 2043–2061, 2020.
[105] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, and D. Xu, "Dependency exploitation: a unified cnn-rnn approach for visual emotion recognition," in *IJCAI*, 2017, pp. 3595–3601.
[106] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *NeurIPS*, 2015, pp. 262–270.
[107] J. Yang, D. She, Y. Lai, and M.-H. Yang, "Retrieving and classifying affective images via deep metric learning," in *AAAI*, 2018, pp. 491–498.
[108] H. Liu, M. Xu, J. Wang, T. Rao, and I. Burnett, "Improving visual saliency computing with emotion intensity," *IEEE TNNLS*, vol. 27, no. 6, pp. 1201–1213, 2016.
[109] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions," in *AAAI*, 2017, pp. 231–237.
[110] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE TMM*, vol. 20, no. 9, pp. 2513–2525, 2018.
[111] S. Zhao, Z. Jia, H. Chen, L. Li, G. Ding, and K. Keutzer, "Pdanet: Polarity-consistent deep attention network for fine-grained visual emotion regression," in *ACM MM*, 2019, pp. 192–201.
[112] T. Rao, X. Li, H. Zhang, and M. Xu, "Multi-level region-based convolutional neural network for image emotion classification," *Neurocomputing*, vol. 333, pp. 429–439, 2019.
[113] X. Yao, D. She, S. Zhao, J. Liang, Y.-K. Lai, and J. Yang, "Attention-aware polarity sensitive embedding for affective image retrieval," in *ICCV*, 2019, pp. 1140–1150.
[114] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[115] R. J. Compton, "The interface between emotion and attention: A review of evidence from psychology and neuroscience," *BCN*, vol. 2, no. 2, pp. 115–129, 2003.
[116] X. Liu, N. Li, and Y. Xia, "Affective image classification by jointly using interpretable art features and semantic annotations," *JVCIR*, vol. 58, pp. 576–588, 2019.
[117] H. Zhang, Z. Yang, M. Gönen, M. Koskela, J. Laaksonen, T. Honkela, and E. Oja, "Affective abstract image classification and retrieval using multiple kernel learning," in *ICONIP*, 2013, pp. 166–175.
[118] X. Lu, N. Sawant, M. G. Newman, R. B. Adams, J. Z. Wang, and J. Li, "Identifying emotions aroused from paintings," in *ECCV*, 2016, pp. 48–63.
[119] S. Zhao, H. Yao, Y. Yang, and Y. Zhang, "Affective image retrieval via multi-graph learning," in *ACM MM*, 2014, pp. 1025–1028.
[120] S. Jindal and S. Singh, "Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning," in *IPSN*, 2015, pp. 447–451.
[121] V. Campos, A. Salvador, X. Giro-i Nieto, and B. Jou, "Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction," in *ASM*, 2015, pp. 57–62.
[122] W. Zhang, X. He, and W. Lu, "Exploring discriminative representations for image emotion recognition with cnns," *IEEE TMM*, vol. 22, no. 2, pp. 515–523, 2020.
[123] V. Campos, B. Jou, and X. Giro-i Nieto, "From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction," *IVC*, vol. 65, pp. 15–22, 2017.
[124] J. Wang, J. Fu, Y. Xu, and T. Mei, "Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks," in *IJCAI*, 2016, pp. 3484–3490.
[125] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.
[126] X. He and W. Zhang, "Emotion recognition by assisted learning with convolutional neural networks," *Neurocomputing*, vol. 291, pp. 187–194, 2018.
[127] M. O. Cordel, S. Fan, Z. Shen, and M. S. Kankanhalli, "Emotion-aware human attention prediction," in *CVPR*, 2019, pp. 4026–4035.
[128] M. Sun, J. Yang, K. Wang, and H. Shen, "Discovering affective regions in deep convolutional neural networks for visual sentiment prediction," in *ICME*, 2016, pp. 1–6.
[129] L. Wu, M. Qi, M. Jian, and H. Zhang, "Visual sentiment analysis by combining global and local information," *NPL*, pp. 1–13, 2019.
[130] S. Fan, M. Jiang, Z. Shen, B. L. Koenig, M. S. Kankanhalli, and Q. Zhao, "The role of visual attention in sentiment prediction," in *ACM MM*, 2017, pp. 217–225.
[131] K. Song, T. Yao, Q. Ling, and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing*, vol. 312, pp. 218–228, 2018.
[132] J. Zhang, M. Chen, H. Sun, D. Li, and Z. Wang, "Object semantics sentiment correlation analysis enhanced image sentiment classification," *KBS*, vol. 191, p. 105245, 2020.
[133] X. He, H. Zhang, N. Li, L. Feng, and F. Zheng, "A multi-attentive pyramidal model for visual sentiment analysis," in *IJCNN*, 2019, pp. 1–8.
[134] J. Yu, C. Cui, L. Geng, Y. Ma, and Y. Yin, "Towards unified aesthetics and emotion prediction in images," in *ICIP*, 2019, pp. 2526–2530.
[135] J. Zhang, H. Sun, Z. Wang, and T. Ruan, "Another dimension: Towards multi-subnet neural network for image sentiment analysis," in *ICME*, 2019, pp. 1126–1131.
[136] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *NIPS*, pp. 1097–1105, 2012.
[137] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.
[138] T. Rui, P. Cui, and W. Zhu, "Joint user-interest and social-influence emotion prediction for individuals," *Neurocomputing*, vol. 230, pp. 66–76, 2017.
[139] S. Zhao, H. Yao, Y. Gao, G. Ding, and T.-S. Chua, "Predicting personalized image emotion perceptions in social networks," *IEEE TAFFC*, vol. 9, no. 4, pp. 526–540, 2018.
[140] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE TPAMI*, vol. 35, no. 10, pp. 2401–2412, 2013.
[141] S. Zhao, G. Ding, Y. Gao, X. Zhao, Y. Tang, J. Han, H. Yao, and Q. Huang, "Discrete probability distribution prediction of image emotions with shared sparse learning," *IEEE TAFFC*, 2018.
[142] S. Zhao, G. Ding, Y. Gao, and J. Han, "Learning visual emotion distributions via multi-modal features fusion," in *ACM MM*, 2017, pp. 369–377.
[143] J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network," in *IJCAI*, 2017, pp. 3266–3272.
[144] T. He and X. Jin, "Image emotion distribution learning with graph convolutional networks," in *ICMR*, 2019, pp. 382–390.
[145] A. Liu, Y. Shi, P. Jing, J. Liu, and Y. Su, "Structured low-rank inverse-covariance estimation for visual sentiment distribution prediction," *SP*, vol. 152, pp. 206–216, 2018.
[146] H. Xiong, H. Liu, B. Zhong, and Y. Fu, "Structured and sparse annotations for image emotion distribution learning," in *AAAI*, 2019.
[147] L. Wang, X. Xu, F. Liu, X. Xing, B. Cai, and W. Lu, "Robust emotion navigation: Few-shot visual sentiment analysis by auxiliary noisy data," in *ACII Workshops and Demos*, 2019, pp. 121–127.
[148] Y. Wang, S. Wang, J. Tang, H. Liu, and B. Li, "Unsupervised sentiment analysis for social media images," in *IJCAI*, 2015, pp. 2378–2379.
[149] L. Wu, S. Liu, M. Jian, J. Luo, X. Zhang, and M. Qi, "Reducing noisy labels in weakly labeled data for visual sentiment analysis," in *ICIP*, 2017, pp. 1322–1326.
[150] F. Chen, R. Ji, J. Su, D. Cao, and Y. Gao, "Predicting microblog sentiments via weakly supervised multimodal deep learning," *IEEE TMM*, vol. 20, no. 4, pp. 997–1007, 2018.

[151] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014, pp. 2672–2680.
[152] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232.
[153] S. Zhao, B. Li, P. Xu, X. Yue, G. Ding, and K. Keutzer, "Madan: multi-source adversarial domain aggregation network for domain adaptation," *IJCV*, 2021.
[154] Y. He and G. Ding, "Deep transfer learning for image emotion analysis: Reducing marginal and joint distribution discrepancies together," *NPL*, pp. 1–10, 2019.
[155] C. Lin, S. Zhao, L. Meng, and T.-S. Chua, "Multi-source domain adaptation for visual sentiment classification," in *AAAI*, 2020.
[156] Q.-T. Truong and H. W. Lauw, "Visual sentiment analysis for review images with item-oriented and user-oriented cnn," in *ACM MM*, 2017, pp. 1274–1282.
[157] J. Ye, X. Peng, Y. Qiao, H. Xing, J. Li, and R. Ji, "Visual-textual sentiment analysis in product reviews," in *ICIP*, 2019, pp. 869–873.
[158] S. Z. Hassan, K. Ahmad, A. Al-Fuqaha, and N. Conci, "Sentiment analysis from images of natural disasters," in *CIAP*, 2019, pp. 104–113.
[159] S. C. Guntuku, D. Preotiuc-Pietro, J. C. Eichstaedt, and L. H. Ungar, "What twitter profile and posted images reveal about depression and anxiety," in *AAAI*, 2019.
[160] H. Lin, J. Jia, Q. Guo, Y. Xue, J. Huang, L. Cai, and L. Feng, "Psychological stress detection from cross-media microblog data using deep sparse neural network," in *ICME*, 2014, pp. 1–6.
[161] S. Bao, H. Ma, and W. Li, "Thupis: A new affective image system for psychological analysis," in *BIBM*, 2014, pp. 1–4.
[162] E. Helmes and J. R. Reddon, "A perspective on developments in assessing psychopathology: a critical review of the mmpi and mmpi-2." *Psychological Bulletin*, vol. 113, no. 3, p. 453, 1993.
[163] N. Garg, B. Wansink, and J. J. Inman, "The influence of incidental affect on consumers' food intake," *Journal of Marketing*, vol. 71, no. 1, pp. 194–206, 2007.
[164] M. B. Holbrook and J. O'Shaughnessy, "The role of emotion in advertising," *Psychology & Marketing*, vol. 1, no. 2, pp. 45–64, 1984.
[165] K. Poels and S. Dewitte, "How to capture the heart? reviewing 20 years of emotion measurement in advertising," *JAR*, vol. 46, no. 1, pp. 18–37, 2006.
[166] S. Hosany and D. Gilbert, "Measuring tourists' emotional experiences toward hedonic holiday destinations," *Journal of Travel Research*, vol. 49, no. 4, pp. 513–526, 2010.
[167] S. Pan, J. Lee, and H. Tsai, "Travel photos: Motivations, image dimensions, and affective qualities of places," *Tourism Management*, vol. 40, pp. 59–69, 2014.
[168] M. Toyama and Y. Yamada, "Categorization of destinations based on tourists' emotional responses," in *TTRA International Conference*, 2013.
[169] S. Hosany and G. Prayag, "Patterns of tourists' emotional responses, satisfaction, and intention to recommend," *Journal of Business Research*, vol. 66, no. 6, pp. 730–737, 2013.
[170] E. S.-H. Tan, "Entertainment is emotion: The functional architecture of the entertainment experience," *Media Psychology*, vol. 11, no. 1, pp. 28–51, 2008.
[171] B. Xing, K. Zhang, S. Sun, L. Zhang, Z. Gao, J. Wang, and S. Chen, "Emotion-driven chinese folk music-image retrieval based on de-svm," *Neurocomputing*, vol. 148, pp. 619–627, 2015.
[172] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, and S.-F. Chang, "Object-based visual sentiment concept analysis and application," in *ACM MM*, 2014, pp. 367–376.
[173] S. Zhao, Y. Li, X. Yao, W. Nie, P. Xu, J. Yang, and K. Keutzer, "Emotion-based end-to-end matching between image and music in valence-arousal space," in *ACM MM*, 2020, pp. 2945–2954.
[174] H. A. Ahmad, S. Koyama, and H. Hibino, "Emotion as a key role in successful acceptance of japanese manga by indonesian readers," in *KEER*, 2012.
[175] Q. You, L. Cao, H. Jin, and J. Luo, "Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks," in *ACM MM*, 2016, pp. 1008–1017.
[176] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *INF*, vol. 49, pp. 69–78, 2019.
[177] X. Wang, J. Jia, J. Tang, B. Wu, L. Cai, and L. Xie, "Modeling emotion influence in image social networks," *IEEE TAFFC*, vol. 6, no. 3, pp. 286–297, 2015.
[178] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.
[179] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *AAAI*, 2019, pp. 3558–3565.
[180] B. Wu, J. Jia, Y. Yang, P. Zhao, and J. Tang, "Understanding the emotions behind social images: Inferring with user demographics," in *ICME*, 2015, pp. 1–6.
[181] B. Wu, J. Jia, Y. Yang, P. Zhao, J. Tang, and Q. Tian, "Inferring emotional tags from social images with user demographics," *IEEE TMM*, vol. 19, no. 7, pp. 1670–1684, 2017.
[182] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *ICML*, 2013, pp. 10–18.
[183] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IROS*, 2017, pp. 23–30.
[184] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "Emotiw 2018: Audio-video, student engagement and group-level affect prediction," in *ICMI*, 2018, pp. 653–656.
[185] X. Guo, B. Zhu, L. F. Polanía, C. Boncelet, and K. E. Barner, "Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions," in *ICMI*, 2018, pp. 635–639.
[186] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE TAFFC*, vol. 3, no. 1, pp. 42–55, 2011.
[187] H. Joho, J. Staiano, N. Sebe, and J. M. Jose, "Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents," *MTA*, vol. 51, no. 2, pp. 505–523, 2011.
[188] S. Zhao, H. Yao, and X. Sun, "Video classification and recommendation based on affective analysis of viewers," *Neurocomputing*, vol. 119, pp. 101–110, 2013.
[189] S. Koelstra and I. Patras, "Fusion of facial expressions and eeg for implicit affective tagging," *IVC*, vol. 31, no. 2, pp. 164–174, 2013.
[190] X. Wang, J. Jia, and L. Cai, "Affective image adjustment with a single word," *TVC*, vol. 29, no. 11, pp. 1121–1133, 2013.

Sicheng Zhao (SM'19) received the Ph.D. degree from Harbin Institute of Technology, Harbin, China, in 2016. He worked as a Research Fellow at Tsinghua University from 2016 to 2017 and at University of California, Berkeley from 2017 to 2020. His research interests include affective computing, multimedia, and computer vision.

Xingxu Yao received the Master degree from Nankai university, Tianjin, China, in 2021. His research interests include computer vision, affective computing, and multimedia.

Guoli Jia will work toward the Master's degree at the College of Computer Science, Nankai University, Tianjin, China. His research interests include computer vision and pattern recognition.

Jufeng Yang received the Ph.D. degree from Nankai University, Tianjin, China, in 2009. He is currently a full professor in the Department of Computer Science, Nankai University and was a visiting scholar with the Vision and Learning Lab, University of California, Merced, USA, from 2015 to 2016. His recent interests include affective computing, image retrieval, fine-grained classification, and medical image recognition.

Guiguang Ding received his Ph.D. degree from Xidian University, China, in 2004. He is currently an associate professor of School of Software, Tsinghua University. His current research centers on the area of multimedia information retrieval, computer vision and machine learning.

Tat-Seng Chua joined the National University of Singapore, Singapore, in 1983, and spent three years as a Research Staff Member with the Institute of Systems Science, National University of Singapore. He was the Acting and Founding Dean of the School of Computing, National University of Singapore, from 1998 to 2000. He is currently the KITHCT Chair Professor with the School of Computing, National University of Singapore. His research interests include multimedia information retrieval, multimedia question answering, and the analysis and structuring of user-generated contents.

Björn W. Schuller heads Imperial College London's Group on Language Audio & Music (GLAM), is CEO of the Audio Intelligence company audEERING, and a Full Professor at University of Augsburg/Germany in Computer Science. He received his diploma, doctoral, and habilitation degrees from TUM in Munich/Germany in EE/IT. Previous positions of his include Full Professor at the University of Passau/Germany and Visiting Professor, Associate, and Scientist at VGTU/Lithuania, University of Geneva/Switzerland, Joanneum Research/Austria, Marche Polytechnic University/Italy, and CNRS-LIMSI/France. His technical publications focus on machine intelligence for affective multimedia analysis. He is a Fellow of the IEEE, President-Emeritus of the AAAC, and the Editor in Chief of the IEEE TAFFC.

Kurt Keutzer (F'96) received his Ph.D. degree in Computer Science from Indiana University in 1984 and then joined the research division of AT&T Bell Laboratories. In 1991 he joined Synopsys, Inc. where he ultimately became Chief Technical Officer and Senior Vice-President of Research. In 1998, Kurt became Professor of Electrical Engineering and Computer Science at the University of California at Berkeley. Kurt's research group is currently focused on using parallelism to accelerate the training and deployment of Deep Neural Networks for applications in computer vision, speech recognition, multi-media analysis, and computational finance. Kurt is a Life Fellow of the IEEE.