Affective Image Content Analysis: A Comprehensive Survey∗

Sicheng Zhao†, Guiguang Ding‡, Qingming Huang§, Tat-Seng Chua¶, Björn W. Schuller∥ and Kurt Keutzer†
†Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA
‡School of Software, Tsinghua University, China
§School of Computer and Control Engineering, University of Chinese Academy of Sciences, China
¶Department of Computing, National University of Singapore, Singapore
∥Department of Computing, Imperial College London, UK

Abstract
Images can convey rich semantics and induce strong emotions in viewers. Recently, with the explosive growth of visual data, extensive research efforts have been dedicated to affective image content analysis (AICA). In this paper, we review the state-of-the-art methods comprehensively with respect to two main challenges – affective gap and perception subjectivity. We begin with an introduction to the key emotion representation models that have been widely employed in AICA. Available existing datasets for performing evaluation are briefly described. We then summarize and compare the representative approaches on emotion feature extraction, personalized emotion prediction, and emotion distribution learning. Finally, we discuss some future research directions.

1 Introduction
With the rapid development of photography technology and the wide popularity of social networks, people have become used to sharing their experiences and expressing their opinions online using images and videos together with text. This gives rise to a huge volume of multimedia data, which results in urgent demand of processing and understanding the visual content. Compared to low-level visual appearances, humans could better perceive and understand the high cognitive level and affective level of semantics [Zhao et al., 2014a]. Existing works on image content analysis from the computer vision community mainly focus on understanding the cognitive aspects of images, such as object detection. Recently, a lot of research attention has been paid from the psychology, affective computing and multimedia communities to affective level analysis of image content. As what people feel may directly determine their decision making, affective image content analysis (AICA) is of great importance, which can enable wide applications [Chen et al., 2014], ranging from human-computer interaction to image retrieval.

‡Corresponding author: Sicheng Zhao (schzhao@gmail.com)

Specifically, 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 [Zhao et al., 2017a]. One main challenge for AICA is the affective gap, 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” [Hanjalic, 2006]. Recently, various hand-crafted or learning-based features have been designed to bridge this gap. Current AICA methods mainly assign an image with the dominant (average) emotion category (DEC) with the assumption that different viewers react similarly to the same image. This task can be performed as a traditional single-label learning problem.

However, labeling the emotions in ground-truth generation is in fact highly inconsistent. Different viewers may have totally different emotional reactions to the same image, which is caused by many personal and situational factors, such as the cultural background, personality and social context [Peng et al., 2015; Zhao et al., 2016; Yang et al., 2017b]. This phenomenon causes the so-called subjective perception problem, as shown in Figure 1. In such cases, just predicting the DEC is insufficient for this highly subjective variable.

To tackle the subjectivity issue, we can conduct two kinds of AICA tasks [Zhao et al., 2016]: 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 associates one instance with multiple emotion labels. However, the importance or extent of different labels is in fact unequal. In such cases, emotion distribution learning would make more sense, which aims to learn the degree to which each emotion describes the instance [Yang et al., 2017b].

In this paper, we concentrate on reviewing the state-of-the-art methods on AICA and outlining research trends. First, we introduce the widely-used emotion representation models and the available datasets for performing AICA evaluation. Second, we summarize and compare the representative approaches on emotion feature extraction, personalized emotion prediction, and emotion distribution learning, corresponding to the affective gap and perception subjectivity challenges. Finally, we discuss potential research directions to pursue.
2 Emotion Models from Cognitive Science Community

Psychologists mainly employ two typical models to represent emotions: categorical emotion states (CES), and dimensional emotion space (DES). CES models consider emotions to be one of a few basic categories, such as fear and joy, etc. Some typical models include Ekman’s six basic emotions [Ekman, 1992] and Mikels’ eight emotions [Mikels et al., 2005]. Specifically, emotions can be classified into positive and negative [Zhao et al., 2018b], sometimes including neutral. In such case, “emotion” is usually called “sentiment”. DES models usually employ a 3D or 2D Cartesian space to represent emotions, such as valence-arousal-dominance (VAD) [Schlosberg, 1954] and activity-temperature-weight [Lee and Park, 2011]. VAD is the most widely used DES, where valence represents the pleasantness ranging from happy to unhappy, arousal represents the intensity of emotion ranging from excited to calm, while dominance represents the degree of control ranging from controlled to in control. Dominance is difficult to measure and is often omitted, leading to the commonly used two dimensional VA space [Hanjalic, 2006], where V is easier to recognize in AICA than A.

The relationship between CES and DES and the transformation from one to the other are studied in [Sun et al., 2009]. For example, positive valence relates to a happy state, while negative valence relates to a sad or angry state. CES is easier for users to understand and label, but may not reflect the subtlety and complexity of emotions. DES is more flexible and richer in descriptive power, but absolute continuous values are difficult for users to distinguish and may be unmeaningful because of the lack of agreed-upon standards for subjective emotion rating. CES is mainly used in classification task, while DES is often employed for regression. If discretized into several constant scores, DES can also be used for classification [Lee and Park, 2011]. Ranking based labeling can be applied to ease DEC comprehension difficulties in raters.

Expected emotion is the emotion that the image creator intends to make people feel, while perceived emotion is the actual emotion that is perceived by the image viewers. Discussing the difference or correlation of various emotion models is out of the scope of this paper. We just list some typical emotion models that have been widely used in AICA, as shown in Table 1.

<table>
<thead>
<tr>
<th>Model</th>
<th>Reference</th>
<th>Type</th>
<th>Emotion states/dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ekman</td>
<td>[Ekman, 1992]</td>
<td>CES</td>
<td>happiness, sadness, anger, disgust, fear, surprise</td>
</tr>
<tr>
<td>Mikels</td>
<td>[Mikels et al., 2005]</td>
<td>CES</td>
<td>amusement, anger, awe, contentment, disgust, excitement, fear, sadness</td>
</tr>
<tr>
<td>Plutchik</td>
<td>[Plutchik, 1980]</td>
<td>CES</td>
<td>amusement, anger, fear, disgust, surprise, trust, joy, sadness</td>
</tr>
<tr>
<td>Sentiment</td>
<td>[Schlosberg, 1954]</td>
<td>DES</td>
<td>mood (VA/D)</td>
</tr>
<tr>
<td>ATW</td>
<td>[Lee and Park, 2011]</td>
<td>DES</td>
<td>activity-temperature-weight</td>
</tr>
</tbody>
</table>

Table 1: Representative emotion models employed in AICA.

3 Datasets

The early datasets for AICA mainly come from the psychology community with small-scale images. Recent large-scale datasets are constructed using images from social networks. The released datasets are summarized in Table 2.

The International Affective Picture System (IAPS) is an emotion evoking image set in psychology [Lang et al., 1997]. It consists of 1,182 documentary-style natural color images depicting complex scenes, such as portraits, babies, animals, landscapes, etc. Each image is associated with an empirically derived mean and standard deviation (STD) of VAD ratings in a 9-point rating scale by about 100 college students. The (IAPSa) dataset includes 246 images [Mikels et al., 2005] selected from IAPS, which are labeled by 20 undergraduate participants with a mean age 19.55 years.

The (Abstract) includes 279 peer rated abstract paintings without contextual content, while the Artistic database (ArtPhoto) consists of 806 artistic photographs from a photo sharing site [Machajdik and Hanbury, 2010], the emotions of which are determined by the artist who uploaded the photo. The IAPSa, ArtPhoto, and Abstract datasets are categorized into eight discrete categories [Mikels et al., 2005]. Each image is labeled with only one emotion in ArtPhoto, the images in IAPSa may be labeled with more than one emotion, while the detailed votes of all emotions are provided for the images in Abstract, from which one can obtain the DEC and emotion distributions. Totally, 395 emotions are labeled on the 246 images in IAPSa. 228 Abstract images are usually used for emotion classification after DEC selection.

The Geneva affective picture database (GAPED) is composed of 520 negative images, 121 positive images and 89 neutral images [Dan-Glauser and Scherer, 2011], labeled by 60 participants with a mean age of 24 years (ranging from 19
Table 2: Released and freely available datasets for AICA, where ‘# Images’ and ‘# Annotators’ respectively represent the total number of images and annotators (f: female, m: male), ‘E, P, B’ are short for expected emotion, perceived emotion and both, respectively.

to 43, STD = 5.9). Besides, these images are also rated with valence and arousal values, ranging from 0 to 100 points. The MART dataset is a collection of 500 abstract paintings from the Museum of Modern and Contemporary Art of Trento and Rovereto [Alameda-Pineda et al., 2016], which were realized by professional artists. The devArt dataset is a collection of 500 amateur abstract paintings obtained from the “DeviantArt” online social network [Alameda-Pineda et al., 2016], one of the largest online art communities. Using the relative score method in [Sartori et al., 2015], these abstract paintings are labeled as positive or negative sentiment.

The Twitter dataset (Tweet) includes 470 positive tweets and 133 negative tweets [Borth et al., 2013]. The FlickrCC dataset is constructed by retrieving the Flickr creative common (CC) images for the 3,000 adjective noun pairs (ANPs) [Borth et al., 2013]. By excluding the images that do not contain the ANP string in the title, tag or description, about 500k Flickr CC images in total are generated for 1,553 ANPs. These images are then mapped to the Plutchik’s Wheel of Emotions with 8 basic emotions, each with 3 scales, such as ecstasy—joy—serenity.

The Flickr dataset consists of 301,903 images based on the Ekman emotion model [Yang et al., 2014]. A word list for each of the six emotion categories is manually defined based on WordNet and HowNet. The emotion category whose word list has the most same words as the adjective words of an image’s tags and comments is assigned to the image.

The original FI dataset consists of 90,000 noisy images collected from Flickr and Instagram by searching the emotion keywords [You et al., 2016b]. The weakly labeled images are further labeled by 225 Amazon Mechanical Turk (AMT) workers, which are selected through a qualification test. The 23,308 images that receive at least three votes from their assigned 5 AMT workers are kept. The number of images in each Ekman emotion category is larger than 1,000.

The Emotion6 dataset [Peng et al., 2015] consists of 1,980 images collected from Flickr by using the emotion keywords and synonyms as search terms. There are 330 images for each emotion category. AMT workers were invited to label the images into the Ekman’s 6 emotions and neutral to obtain the emotional responses. Each image was scored by 15 subjects. The discrete emotion distribution information is released.

The IESN dataset is constructed for personalized emotion prediction [Zhao et al., 2016], with 1,012,901 images from Flickr. Lexicon-based methods are used to segment the text of metadata from uploaders for expected emotions and comments from viewers for actual emotions. Synonym based searching is employed to obtain the Ekman’s emotion category by selecting the most frequent synonyms. The average VAD values of the segmentation results are computed as DES ground truth based on the VAD norms of 13,915 English lemmas [Warriner et al., 2013]. 7,723 active users with more than 50 involved images are selected. The DEC and emotion distributions can also be easily obtained.

Two image datasets for discrete emotion distribution learning are released in [Yang et al., 2017b]. One is FlickrLDL dataset, a subset of FlickrCC. FlickrLDL contains 10,700 images, which are labeled by 11 viewers using Mikels’ emotion model. 30,000 images are collected by searching various sentiment key words from Twitter. After duplication removal, the images are labeled by 8 viewers. In this way, the TwitterLDL dataset is generated with 10,045 images. In both datasets, the ground truth emotion distribution for each image is obtained by integrating the votes from the workers.

4 Emotion Feature Extraction

As shown in [Zhao et al., 2014c], there are various types of features that may contribute to the perception of image emotions. In this section, we introduce the hand-crafted features of different levels (Table 3) and the learning based features (Table 5) that have been widely extracted for AICA.

4.1 Low-level Features

Low-level features suffer from being difficult to be interpreted by humans. The widely used features include GIST, HOG2x2, self-similarity and geometric context color histogram features as in [Patterson and Hays, 2012], because they are each individually powerful and can describe distinct visual phenomena in a scene perspective.

Features derived from elements of art, including color and texture are extracted [Machajdik and Hanbury, 2010]. Lee and Park [2011] used the MPEG-7 visual descriptors, including four color-related ideas and two texture-related ideas. Lu et al. [2012] investigated how shape features in natural images influence emotions by modeling the concepts of
4.2 Mid-level Features

Mid-level features are more interpretable, more semantic and more relevant to emotions than low-level features. 102 attributes are detected [Patterson and Hays, 2012], including 5 different types: materials, surface properties, functions or affordances, spatial envelope attributes and object presence. Besides the 102 attributes, Yuan et al. [2013] also incorporated eigenfaces corresponding to different emotions, which contribute a lot to the images containing faces. Rao et al. [2016a] extracted SIFT features as basic feature and adopted bag-of-visual-words (BoVW) to represent the multi-scale blocks. The latent topic distribution estimated by probabilistic latent semantic analysis is used as another mid-level representation.

Harmonious composition is essential in an artwork and Machajdik and Hanbury [2010] extracted simple semantic concepts from principles of art are designed in [Zhao et al., 2014a]. Itten’s color wheel, Sartori et al. [2015] designed two types of features to represent the color combinations.

4.3 High-level Features

High-level features are the semantic contents contained in images. People can easily understand the emotions conveyed in images by recognizing the semantics. In the early years, Machajdik and Hanbury [2010] extracted simple semantic content by detecting faces and skins contained in an image. Facial expressions may determine the emotion of the images containing faces. 8 kinds of facial expressions are extracted as high-level features [Yang et al., 2010]. The expressions of images detected without faces are set as neutral. An 8 dimensional vector, each element of which indicates the number of related facial expressions in the image, is generated.

Borth et al. [2013] proposed to describe the semantic concepts by 1,200 adjective noun pairs (ANPs), which are detected by SentiBank. The advantages of ANP are that it turns potentially cleaner subset of the training instances.

With the advent of deep learning, emphasis has been shifted from designing hand-crafted features to learning features in an end-to-end fashion. To tackle the weakly labeled images, You et al. [2015] proposed to progressively select a potentially cleaner subset of the training instances.
Table 5: Representative works on deep learning based AICA methods, where ‘Pre’ indicates whether the network is pre-trained using ImageNet, ‘# Feat’ indicates the dimension of last feature mapping layer before the emotion output layer, ‘Cla’ indicates the classifier used after the last feature mapping with default Softmax, ‘Loss’ indicates the loss objectives besides the common cross-entropy loss, and ‘Result’ is the reported best accuracy for classification and discounted cumulative gain for retrieval on the corresponding datasets.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Base net</th>
<th>Pre</th>
<th># Feat</th>
<th>Cla</th>
<th>Loss</th>
<th>Dataset</th>
<th>Target</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>[You et al., 2015]</td>
<td>self-defined</td>
<td>no</td>
<td>24</td>
<td>–</td>
<td>–</td>
<td>FlickrCC, unreleased Twitter</td>
<td>cla</td>
<td>0.781</td>
</tr>
<tr>
<td>[You et al., 2016b]</td>
<td>AlexNet</td>
<td>yes</td>
<td>4096</td>
<td>SVM</td>
<td>–</td>
<td>FI, IAPSa, Abstract, ArtPhoto</td>
<td>cla</td>
<td>0.583, 0.872, 0.776, 0.737</td>
</tr>
<tr>
<td>[You et al., 2016b]</td>
<td>AlexNet, ACNN, TCNN</td>
<td>yes</td>
<td>4096, 256, 4096</td>
<td>MIL</td>
<td>–</td>
<td>FI, IAPSa, Abstract, ArtPhoto, MAPR</td>
<td>cla</td>
<td>0.652, 0.889, 0.825, 0.834, 0.764</td>
</tr>
<tr>
<td>[Zhu et al., 2017]</td>
<td>self-defined</td>
<td>no</td>
<td>512</td>
<td>–</td>
<td>–</td>
<td>FI, IAPSa, ArtPhoto</td>
<td>cla</td>
<td>0.730, 0.902, 0.855</td>
</tr>
<tr>
<td>[Yang et al., 2018]</td>
<td>GoogleNet-Inception</td>
<td>yes</td>
<td>1024</td>
<td>–</td>
<td>–</td>
<td>FI, IAPSa, Abstract, ArtPhoto</td>
<td>cla, ret</td>
<td>0.676, 0.442, 0.382, 0.400, 0.780, 0.819, 0.788, 0.704</td>
</tr>
</tbody>
</table>

convolutional neural network (CNN) model is trained on the training data. According to the prediction score of the trained model on the training data itself, the training samples with distinct sentiment scores between the two classes with a high probability are selected. The trained model is fine-tuned using the newly selected instances. Later, they fine-tuned the pre-trained AlexNet on ImageNet to classify emotions into 8 categories by changing the last layer of the neural network from 1000 to 8 [You et al., 2016b]. An SVM classifier is also trained using features extracted from the second to the last layer of the pre-trained AlexNet model.

Rao et al. [2016b] proposed to learn multi-level deep representations for image emotion classification (MldrNet). The input image is segmented into 3 levels of patches, which are input to 3 different CNN models, including Alexnet, aesthetics CNN (ACNN) and texture CNN (TCNN). Multiple instance learning (MIL) is employed to generate the emotion label of an input image. Based on MldrNet, Zhu et al. [2017] integrated the different levels of features by a Bidirectional GRU model (BiGRU) to exploit their dependencies. Two features generated from our Bi-GRU model are concatenated as the final features to predict the emotions. Apart from the traditional cross-entropy loss, an additional contrastive loss is jointly optimized to enforce the feature vectors extracted from each pair of images from the same category to be close enough and those from different categories to be far away.

To explore the correlation of emotional labels with the same polarity, Yang et al. [2018] employed deep metric learning and proposed a multi-task deep framework to optimize both retrieval and classification tasks. Besides the cross-entropy loss, a novel sentiment constraint is jointly optimized by considering the relations among emotional categories in the Mikels’ wheel, which extends triplet constraints to a hierarchical structure. A sentiment vector based on the texture information from the convolutional layer is proposed to measure the difference between affective images.

The deep representations generally outperform the hand-crafted features, which are designed based on several small-scale datasets for specific domains. However, how the deep features correlate to specific emotions is unclear.
Table 6: Representative discrete distribution learning works on AICA, where ‘Fusion’ indicates the fusion strategy of different features, ‘Pre’ indicates whether the network is pre-trained using ImageNet, ‘# Output’ is the output dimension of the last layer, and ‘Result’ is the reported best KL divergence on the corresponding datasets except the first line, which is the result on sum of squared difference.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Feature</th>
<th>Fusion</th>
<th>Learning model</th>
<th>Dataset</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Zhao et al., 2015]</td>
<td>GIST, Elements, Principles</td>
<td>-</td>
<td>SSL</td>
<td>Abstract</td>
<td>0.134</td>
</tr>
<tr>
<td>[Zhao et al., 2017a]</td>
<td>GIST, Elements, Attributes, Principles, ANP, deep features from AlexNet</td>
<td>-</td>
<td>WMMSSL</td>
<td>Abstract, Emotion6, IESN</td>
<td>0.482, 0.479, 0.478</td>
</tr>
<tr>
<td>[Yang et al., 2017b]</td>
<td>ANP, deep features from VGG16</td>
<td>-</td>
<td>augmented CPNN</td>
<td>Abstract, Emotion6, FlickrLDL, TwitterLDL</td>
<td>0.480, 0.506, 0.469, 0.555</td>
</tr>
<tr>
<td>[Zhao et al., 2017b]</td>
<td>GIST, Elements, Attributes, Principles, ANP, deep features from AlexNet</td>
<td>-</td>
<td>WMMC-PNN</td>
<td>Abstract, Emotion6, IESN</td>
<td>0.461, 0.464, 0.470</td>
</tr>
</tbody>
</table>

Figure 2: Examples of the image emotion’s discrete probability distribution [Zhao et al., 2018a].

Both SSL and WMMSSL can only model one test image each time. For different test images, the shared coefficients have to be learned repeatedly. Geng et al. [2013] proposed a conditional probability neural network (CPNN) for distribution learning by modeling the conditional probability of labels given features as a three layer neural network. Yang et al. [2017b] replaced the signless integers in CPNN with a binary representation, since adding two emotion labels or subtracting one label from another is meaningless. By adding noises to the ground truth emotion labels, more roughly labeled distributions and samples are generated. Zhao et al. [2017b] extended CPNN into multi-modal settings and proposed weighted multi-modal CPNN (WMMC-PNN) to jointly explore the representation abilities of different features. The linear combination of different CPNN loss functions are minimized with a sparse constraint on the combination coefficients. Once the parameters of the augmented CPNN and WMMC-PNN are learned, the emotion distributions of a test image would be easily obtained.

The CPNN series have only 3 layers and the input is extracted visual features instead of original images. A deeper CNN regressor (CNNR) for each emotion category is trained in Emotion6 [Peng et al., 2015] with the exact AlexNet architecture. The number of output nodes is changed to 1 to predict a real value and the Softmax loss layer is replaced with the Euclidean loss layer. The predicted probabilities of all emotion categories are normalized to sum to 1. However, the CNNR cannot guarantee that the predicted probability is non-negative. Further, the probability correlations among different emotions are actually ignored, since the regressor for each emotion category is trained independently. Yang et al. [2017a] designed a multi-task deep framework based on VGG16 by jointly optimizing the cross-entropy loss for emotion classification and Kullback-Leibler (KL) divergence loss for emotion distribution learning, which achieves the state-of-the-art performances. For the single emotion dataset, the single label is transformed to emotion distribution with emotion distances computed on Mikels’ wheel. The representative methods are summarize in Table 6.

6.2 Continuous Probability Distribution Learning
Zhao et al. [2017c] proposed to learn continuous emotion distributions in V A space. Based on the assumption that the VA emotion labels can be well modeled by a mixture of 2 bidimensional Gaussian mixture models (GMMs, see Figure 1(e)), the task turns to predict the parameters of GMMs, including the mean vector and covariance matrix of the 2 Gaussian components as well as the mixing coefficients.

Shared sparse regression (SSR) is proposed 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. SSR can only model one test image each time. To explore the task relatedness, multi-task SSR is further presented to simultaneously predict the parameters of different test images by utilizing appropriate shared information across tasks. GIST, Elements, Attributes, Principles, ANP, and 4,096-dimensional deep features from AlexNet are extracted as visual features. Experiments are conducted on a subset of IESN, which consists of 18,700 images each with more than 20 VA labels. The average KL divergence of multi-task SSR using ANP is 0.436.

7 Conclusion and Future Directions
This paper attempted to provide an overview of recent developments on affective image content analysis (AICA). Obviously, it cannot cover all the literature on AICA, and we focused on a representative subset of the latest methods. We summarized the widely employed emotion representation models, released datasets, and compared the existing methods on emotion feature extraction, personalized emotion prediction and emotion distribution learning. We believe that AICA will continue to be an active and promising research area with broad potential applications, such as an emotion-aware personalized music slide show [Chen et al., 2014], emotion based image musicalization [Zhao et al., 2014b], and image...
captioning with sentiment [Mathews et al., 2016]. Many issues in AICA, however, are still open.

**Understanding image content and context.** Accurately analyzing what is contained in an image can significantly improve the performance of AICA. Sometimes we even need subtle analysis of visual contents. For example, 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”. The context of image content is also important in AICA. Similar visual content under different contexts may evoke different emotions. For example, we may feel “happy” about beautiful flowers. But if the flowers are placed in a funeral, we possibly feel “sad”.

**Recognizing group emotions.** Recognizing the dominant emotion is too generic, while predicting personalized emotion is too specific. Modeling emotions for groups or cliques of users, who share similar tastes or interests, may be a good choice – e.g., by cultural or societal background. Analyzing the user profiles provided by each individual to classify users into different types of backgrounds, tastes and interests may help to tackle this problem.

**Understanding emotions of 3D data and videos.** Compared with traditional intensity images, 3D data contain more spatial information, being useful in low light levels and being color and texture invariant, while videos (such as animated GIF) contain rich temporal correlation information. Combining the spatial and temporal correlation together with the visual content would make more sense.

Further, jointly exploring the complementarity of multimodal data (such as image and text) [You et al., 2016a; Zhao et al., 2018c; Chen et al., 2018] would improve the emotion recognition performance. How to adapt the emotions from one labeled domain to another unlabeled domain [Patel et al., 2015] is another interesting research topic.

**Acknowledgments**

This work is supported by the Berkeley Deep Drive, the National Natural Science Foundation of China (Nos. 61701273, 61571269, 61332016, 61620106009, U1636214), the China Postdoctoral Science Foundation Project (No. 2017M610897), and the European Union’s Horizon 2020 Programme through the Innovative Action No. 645094 (SEWA). This research is part of NExT++ project, supported by the National Research Foundation, Prime Ministers Office, Singapore under its IRC@Singapore Funding Initiative.

**References**


[Zhao et al., 2018b] Sicheng Zhao, Guiguang Ding, Jungong Han, and Yue Gao. Personality-aware personalized emotion recognition from physiological signals. In International Joint Conference on Artificial Intelligence, 2018.

