

EmoSet: A Large-scale Visual Emotion Dataset with Rich Attributes

Jingyuan Yang¹, Qirui Huang¹, Tingting Ding¹, Dani Lischinski², Daniel Cohen-Or³, Hui Huang^{1*}

¹Shenzhen University ²The Hebrew University of Jerusalem ³Tel Aviv University

{jingyuanyang.jyy, qrhuang2021, dingt6616, danix3d, cohenor, hhzhiyan}@gmail.com



Figure 1: EmoSet images are annotated with *eight emotion categories* (blue) and *six emotion attributes* (orange), where different attributes may evoke different emotions.

Abstract

Visual Emotion Analysis (VEA) aims at predicting people’s emotional responses to visual stimuli. This is a promising, yet challenging, task in affective computing, which has drawn increasing attention in recent years. Most of the existing work in this area focuses on feature design, while little attention has been paid to dataset construction. In this work, we introduce EmoSet, the first large-scale visual emotion dataset annotated with rich attributes, which is superior to existing datasets in four aspects: scale, annotation richness, diversity, and data balance. EmoSet comprises 3.3 million images in total, with 118,102 of these images carefully labeled by human annotators, making it five times larger than the largest existing dataset. EmoSet includes images from social networks, as well as artistic images, and it is well balanced between different emotion categories. Motivated by psychological studies, in addition to emotion category, each image is also annotated with a set of describable emotion attributes: brightness, colorfulness, scene type, object class, facial expression, and human action, which can help understand visual emotions in a precise and interpretable way. The relevance of these emotion attributes is validated by analyzing the correlations between them and visual emotion, as well as by designing an attribute module to help visual emotion recognition. We be-

lieve EmoSet will bring some key insights and encourage further research in visual emotion analysis and understanding. Project page: <https://vcc.tech/EmoSet>.

1. Introduction

Emotions are different ways to think that our mind uses to increase our intelligence [31]. Much of the research in Artificial Intelligence (AI) has focused on designing human-like machines, while neglecting emotional intelligence. Since emotions are innate to human beings, AI systems should aim to better understand emotions, in order to succeed in mimicking human behavior. Affective computing [35] is an emerging field that aims to identify, understand, and respond to human emotions. This field has seen significant progress in recent years, and has potential applications in areas such as education [49], healthcare [56], advertising [42], and safety [19].

Visual Emotion Analysis (VEA) is a promising, yet challenging, task in affective computing, aiming to predict emotional responses to visual stimuli. For instance, when viewing the images in Figure 1, one not only recognizes the visual elements therein, but also may experience emotional reactions. Furthermore, even though emotions are subjective, people tend to share similar reactions to the same external stimuli. With the prevalence of social networks, users often choose to convey feelings via images shared on the in-

*Corresponding author.

ternet. Thus, VEA is an increasingly popular research topic within the computer vision field [51, 63]. Advances in VEA may benefit high-level vision tasks (*e.g.*, image aesthetic assessment [26], stylized image captioning [12], and image understanding [47]), as well as human-centered applications (*e.g.*, opinion mining [27], mental health [46], smart advertisement [41], and hate detection [1]).

Most of the work in VEA focused on feature design, covering hand-crafted features [28, 60, 3] and, more recently, learned ones [37, 53, 51]. Based on art and psychological theories, hand-crafted features fail to cover all important factors in human emotions. Although deep learning methods boost recognition performance significantly, the results are still unsatisfying. In particular, while supervised deep learning methods often require large-scale labeled datasets, little attention has been paid to dataset construction. Existing VEA datasets are usually unlabeled on a large scale or labeled on a relatively small scale [24, 58, 34]. Besides, only emotion labels are provided in most datasets. Since emotions are abstract, a key problem is how to bridge the *affective gap* [13] between images and emotions with auxiliary information. We believe that a new and rich dataset is needed for further research and improvement in VEA.

To tackle the above issues, we introduce EmoSet, a large-scale visual emotion dataset, annotated with rich attributes. EmoSet is superior to existing datasets in four aspects: scale, annotation richness, diversity, and data balance. The full (EmoSet-3.3M) dataset comprises 3.3 million machine retrieved and annotated images, among which there are 118,102 human-annotated ones (EmoSet-118K). The latter is five times larger than the widely-used FI dataset [58], as reported in Table 1. Apart from emotions, our dataset is annotated with *emotion attributes*. Inspired by psychological studies [22, 4, 7], we propose a set of describable visual attributes to facilitate understanding *why* an image evokes a certain emotion. Considering the complexity of emotions, the attributes are designed to cover different levels of visual information, including brightness, colorfulness, scene type, object class, facial expression, and human action. With these rich attribute annotations, we hope EmoSet will improve not only the recognition of visual emotions, but also their understanding.

By querying 810 emotion keywords based on Mikels model [29], we collect 3.3 million candidate images from four different sources to form the EmoSet-3.3M dataset. A subset of EmoSet-3.3M is then labeled by human annotators, yielding the EmoSet-118K dataset. Compared with existing datasets, EmoSet contains diverse images covering both social and artistic types. Furthermore, EmoSet-118K is well balanced between the eight emotion categories, each of which is represented with 10,660 to 19,828 images, as reported in Table 2. We further analyze the correlations between our attributes and emotion categories, and demon-

strate that some attributes are indeed strongly relevant to emotions. In addition, to mine the emotion-related information from each attribute, we design an attribute module for visual emotion recognition, and validate it using several CNN backbones.

In summary, our contributions are:

- EmoSet, the first large-scale visual emotion dataset with rich attributes, exceeding existing VEA datasets in terms of scale, annotation richness, diversity and data balance.
- A set of describable emotion attributes motivated by psychological studies, which help understand visual emotional stimuli in a precise and interpretable way.
- A series of in-depth analyses on EmoSet, to explore how emotion attributes advance emotion understanding. Statistical analysis shows that correlations do occur between emotions and attributes, which is consistent with human cognition.
- An attribute module to facilitate visual emotion recognition. Experimental results and visualizations further validate the relevance of emotion attributes and show their potential in understanding visual emotions.

2. Related Work

2.1. Visual Emotion Datasets

In psychology, emotion models can be grouped into two types: Categorical Emotion States (CES), where emotions are described with discrete categories, and Dimensional Emotion Space (DES), where a continuous space is used to represent emotions [62]. CES models are more popular in VEA for their simplicity and interpretability, including 2-category sentiment model, 6-category Ekman model [7], and 8-category Mikels model [29].

Numerous datasets have been constructed to study people’s different emotional reactions toward images [29, 34, 58]. In Table 1 we report various statistics of widely-used VEA datasets, as well as our new dataset, EmoSet. Considering the diversity in image types, emotion models and dataset scales, below we discuss five of the existing datasets.

IAPSa. IAPS [24] aims to investigate the possible relationships between emotions and visual stimuli. IAPSa [29] is a subset of IAPS, built on the Mikels model with eight emotion categories covering amusement, awe, contentment, excitement, anger, disgust, fear, and sadness. IAPSa comprises 395 affective images, and is the first visual emotion dataset with discrete categories.

ArtPhoto. There are a total of 806 artistic images in ArtPhoto [28], which is collected from an art sharing website by using emotion categories as the search keywords.

Emotion6. Emotion6 [34] has 1,980 images collected from Flickr. Each image is labeled by 15 annotators based

Table 1: Comparison between VEA datasets.

Dataset	#Image	#Attribute	#Annotator (per image)	Model (#category)	Image type
IAPSa [29]	395	-	-	Mikels (8)	Natural
Abstract [28]	280	-	14	Mikels (8)	Abstract
ArtPhoto [28]	806	-	-	Mikels (8)	Artistic
Twitter I [57]	1,269	-	5	Sentiment (2)	Social
Twitter II [3]	603	-	3	Sentiment (2)	Social
Emotion6 [34]	1,980	-	15	Ekman (6)	Social
HECO [50]	9,385	-	-	(8)	Social
Flickr [17]	60,745	-	3	Sentiment (2)	Social
Instagram [17]	42,856	-	3	Sentiment (2)	Social
FI [58]	23,308	-	5	Mikels (8)	Social
EMOTIC [20]	18,316	-	3	Ekman (6)	Social
FlickrLDL [55]	10,700	-	11	Mikels (8)	Social
TwitterLDL [55]	10,045	-	8	Mikels (8)	Social
EmoSet	118,102	6	10	Mikels (8)	Social, Artistic

on Ekman model, covering six emotion categories, including happiness, anger, disgust, fear, sadness, and surprise.

Flickr and Instagram. Images in Flickr and Instagram [17] are crawled from the internet by searching emotional categories. After labeling by crowd-sourced human annotation, 60,745 and 42,856 affective images are preserved with sentiment labels (*i.e.*, positive or negative). Specifically, each image is labeled by 3 workers, where the ground-truth is determined by a majority vote.

FI. Another dataset based on Flickr and Instagram is FI [58]. It is one of the largest scale visual emotion datasets to date, with a total of 23,308 labeled images. Using eight emotion words in Mikels model as queries, FI collects candidate images from Flickr and Instagram. Each collected image is then labeled by 5 Amazon Mechanical Turk (AMT) workers, and images having more than 3 votes are kept in the final dataset. FI serves as one of the most widely-used datasets in VEA.

In this work, we introduce EmoSet, which is five times larger than the FI dataset. In addition to the Mikels emotion category, each image is annotated with six comprehensively designed emotion attributes, including brightness, colorfulness, scene type, object class, facial expression and human action. The images in our dataset come from more diverse sources and the dataset is more balanced across emotion categories, as reported in Tables 1 and 2.

2.2. Visual Emotion Recognition

Researchers have been engaged in VEA for two decades, with approaches ranging from the early traditional ones to the recent deep learning ones. Machajdik *et al.* [28] extracts specific image features to predict emotions, *i.e.*, color, texture, composition, and content inspired by psychology and art theory. Adjective Noun Pairs (ANPs) are introduced by Borth *et al.* [3] to help learn visual emotions from a semantic level. By extracting a set of principle-of-art-based emotional features like balance, emphasis, harmony, vari-

ety, gradation, and movement, Zhao *et al.* [60] proposes a method to deal with both classification and regression tasks. Traditional methods fail to cover all important factors related to human emotions, leading to sub-optimal results.

Based on deep learning techniques, You *et al.* [57] propose a progressive CNN (PCNN), and Rao *et al.* [37] build a multi-level deep representation network (MldrNet). Early attempts usually focus on extracting holistic image features, while neglecting the importance of local regions. Yang *et al.* [54] leverages object detection, as well as attention mechanism [53] to help emotion recognition. With specially designed emotional features, Yang *et al.* [52] construct a network to learn emotions from different visual stimuli and to mine the correlations between them [51]. However, development in VEA is still unsatisfying due to low classification accuracy and the use of generic network design. Considering the abstract nature of emotion, it is necessary to introduce auxiliary information to assist visual emotion recognition. We believe that a large-scale dataset with rich annotations can help to mitigate the affective gap [13] between images and emotion labels.

3. Construction of EmoSet

3.1. Data Collection

To build EmoSet, we need to retrieve a large number of images from the Internet. Since not all images are likely to arouse significant emotion, we construct a list of emotion keywords to help filter candidate images for our dataset. Following previous work [28, 58], EmoSet is built on the widely-used Mikels model [29] with eight categories, *i.e.*, amusement, awe, contentment, excitement, anger, disgust, fear, sadness, where the former four are positive emotions and the latter four are negative ones. Each of the eight emotion categories is first synonymized according to three widely-used English dictionaries: WordNet [30], Collins [14] and Roget’s [39]. For instance, “sadness” is synonymized to words like “depression, sorrow, mourn, despair, grieve”. Since the number of retrievable images is often limited for each query, we combine the synonyms and further augment them with different parts of speech, aiming at retrieving a large amount of data. For example, “amusement” is augmented with other word forms like “amuse, amuses, amused, amusing, amusingly”. The final list contains 810 keywords, serving as queries to retrieve candidate images from the Internet. For more details, please refer to the supplementary material.

In view of the fact that in most image-text pairs, the two modalities are in agreement, *i.e.*, the textual tag or description indeed reflects the emotion that the image conveys, each retrieved image is automatically labeled with one of the eight emotions from which the query used to retrieve that image was derived. For larger scale and richer

diversity, EmoSet is collected from four different sources including openverse, pexels, pixabay and rawpixels. In total, 4.3 million images are collected, followed by voting in order to determine the labels of images with more than one emotion tag and removal of duplicates considering both file names and pixel-wise similarities, leaving 3.3 million images. With such a large amount of images tagged with text descriptions, EmoSet-3.3M has great potential for weakly supervised learning [65], vision-language modeling [36] and multi-modal emotion analysis [44].

3.2. Emotion Attributes Design

Aiming at figuring out some possible visual cues related to human emotions, we propose a set of describable emotion attributes inspired by psychological studies. Since emotions are abstract and complex, emotion attributes are designed to cover different levels of visual information: low-level (*i.e.*, brightness, colorfulness), mid-level (*i.e.*, scene type, object class), and high-level (*i.e.*, facial expression, human action). We leverage both classic traditional methods and well-trained deep models to automatically predict attributes, which then constitute part of the automatic annotation, along with the emotion labels.

Brightness: The overall lighting level of an image has been proven to be essential in perceptual processing and is closely related to human emotions [22]. We use discrete numerical values ranging from 0 (darkest) to 1 (brightest), with increments of 0.1, to quantify the brightness of an image [5]. For more details, please refer to the supplementary material.

Colorfulness: Psychological studies [38] suggest that there are correlations between the perceived colorfulness of an image and affect. Similarly to brightness, colorfulness is calculated, normalized and discretized to the range of 0 to 1 [32]. Specifically, 0 corresponds to black-and-white images, while 1 refers to the most colorful ones.

Scene type: The scene depicted in an image is often considered as an important emotional stimulus [4]. We use a scene recognition model trained with Places365 [64], a well-known benchmark for scene recognition. Out of 365 scene categories (*e.g.*, sky, mountain, balcony, plaza, and church), we choose the top prediction as the scene type label for each image in our dataset.

Object class: Psychologists have long investigated the relationships between objects and emotions [10]. Thus, we associate object labels with each image in our dataset. Our object detection model is built on the

<https://wordpress.org/openverse/>
<https://www.pexels.com/>
<https://pixabay.com/>
<https://www.rawpixel.com/>

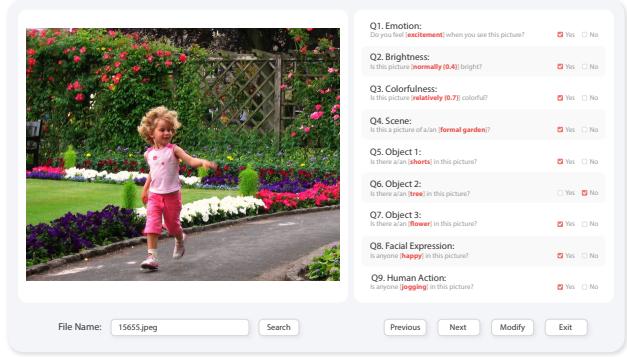


Figure 2: Graphical interface of annotation tool.

OpenImagesV4 dataset [23]. Considering that multiple objects may appear in an image and jointly evoke emotions, we associate with each image the three object classes predicted with the highest confidence.

Facial expression: Facial expression can undoubtedly influence visual emotion experience [7], where people tend to empathize with the one in image. In Ekman model, there are six basic facial expressions: happy, angry, disgust, fear, sad, and surprise. We crop the largest face in the image and apply a model pre-trained on FER2013 [11] to obtain the facial expression label.

Human action: Some human actions stem from emotion and can also arouse emotion in an observer [66]. Kinetics 400 is a large video dataset for human action recognition [18], which includes various actions like dining, water sliding, playing piano, barbecuing, and training a dog. Since Kinetics 400 is based on the video modality, we convert the image into a single-frame video as input and feed it into the UnifornerV2 model [25] to predict the human action label.

3.3. Human Annotation

EmoSet-3.3M is automatically labeled by queries (*i.e.*, emotions) and machines (*i.e.*, attributes) without human participation. To build a more carefully annotated dataset, we invite humans to help annotate and ask them to take the qualification tests first. We ask the participants to take the empathy quotient test [2] to verify that they are sensitive to emotions, where annotators are qualified with a score greater than 30. Subsequently, we randomly select 100 emotion-labeled images from the FI dataset to evaluate the classification accuracy of the participants, with a passing rate at 85%. We hired 60 annotators who passed all the above tests, thereby meeting our criteria.

There are three main challenges in visual emotion analysis: abstractness, ambiguity and subjectivity. For abstractness, we introduce a set of attributes to help understand emotion in a more precise and interpretable way. The annotation tool is presented in Figure 2, where annotators are

Table 2: Comparison on image number per category, where **purple** indicates the maximum while **green** the minimum.

Dataset	Amusement	Anger	Awe	Contentment	Disgust	Excitement	Fear	Sadness	Total
IAPSa	37 (9%)	8 (2%)	54 (14%)	63 (16%)	74 (19%)	55 (14%)	42 (11%)	62 (16%)	395
Arphoto	101 (13%)	77 (10%)	102 (13%)	70 (9%)	70 (9%)	105 (13%)	115 (14%)	166 (21%)	806
Abstract	25 (11%)	3 (1%)	15 (7%)	63 (28%)	18 (8%)	36 (16%)	36 (16%)	32 (14%)	228
FI	4942 (21%)	1266 (5%)	3151 (14%)	5374 (23%)	1658 (7%)	2963 (13%)	1032 (4%)	2922 (13%)	23308
EmoSet	19445 (16%)	10660 (9%)	15037 (13%)	16337 (14%)	10666 (9%)	19828 (17%)	13453 (11%)	12676 (11%)	118102



Figure 3: Word cloud distributions of scene type, object class and human action, where the larger the font, the higher the frequency it appears.

required to answer several questions on emotion (Q1) and attributes (Q2-Q9). For example, annotators are asked “Do you feel *excitement* when you see this picture?” (emotion) or “Is this a picture of *formal garden*?” (attribute). Since emotions are ambiguous, it is much easier for annotators to indicate whether an image evokes a specific emotion, rather than asking them to decide which emotion a given image evokes. Fewer choices may lead to more accurate results. Thus, we ask the annotators to verify both the emotion and the attribute labels for each image by answering “yes” or “no”, instead of selecting a specific category, following previous work [6, 58]. To mitigate the subjectivity in emotion annotations, in EmoSet, each image is labeled by 10 annotators. For each image, annotation results that reached a consensus of more than 7 out of 10 annotators are regarded as the final label. In particular, images with more than 7 votes for “yes” in emotion label are preserved while others are deleted. For more details, please refer to the supplementary material. By the end, EmoSet-118K is carefully labeled with human annotations, where both emotion labels and attribute labels are provided. Note that the analysis and evaluation in this paper are reported on EmoSet-118K.

4. Analysis of EmoSet

4.1. Properties of EmoSet

EmoSet aims at constructing a comprehensive and interpretable dataset, which can help researchers to dive deep into visual emotions. To our knowledge, this is the first large-scale VEA dataset that is also annotated with rich attributes, as shown in Table 1. In general, EmoSet has advantages in four aspects compared with the existing datasets: scale, annotation richness, diversity, and data balance.

Scale: Built with 3.3 million images with 118,102 human annotated ones, EmoSet is five times larger than the existing large-scale dataset FI (23,308). To our knowl-

edge, it is the largest labeled visual emotion dataset in terms of the total number of images as well as the number of images per category.

Annotation richness: Based on Mikels model, EmoSet is labeled with eight emotion categories. In addition to emotion labels, we annotate EmoSet with six emotion attributes under different categories, which include *brightness* (10), *colorfulness* (10), *scene type* (365), *object class* (600), *facial expression* (6), and *human action* (400), to help understand emotions in a fine-grained manner. It is the first VEA dataset with attributes. In Figure 3, we present the word cloud distribution of scene type, object class and human action.

Diversity: Images are queried by 810 emotion keywords from four different sources, sharing a large data discrepancy. Different from the previous single-type emotion datasets, in EmoSet, there are images uploaded by social media users, as well as artistic work shared by professional photographers. For more details, please refer to the supplementary material.

Data balance: As shown in Table 2, existing VEA datasets are unevenly distributed among different emotion categories, with the minimum and maximum numbers indicated by different colors. In the Abstract dataset, for example, the *anger* category is represented by only 1% of the images, while 28% represent *contentment*. Data balance is essential for learning a good recognition model. Thus, our EmoSet is built with a balanced data distribution, where the number of images in each category is between 10,660 and 19,828.

4.2. Emotion-Attribute Analysis

Emotion attributes are designed to help visual emotion recognition as well as understanding. To validate their effectiveness, we conduct several numerical experiments and visualizations on EmoSet to examine the relationships between attributes and emotions.

In Mikels model, amusement, awe, contentment, and excitement are considered *positive* emotions, while anger, disgust, fear, and sadness are *negative* ones. The two polarities (*i.e.*, positive, negative) can be seen as more basic emotional elements, than the eight more specific categories. One might hypothesize that brighter and/or more colorful images are more likely to evoke a positive emotion. We verify this hypothesis in Figure 4, where we plot the break-

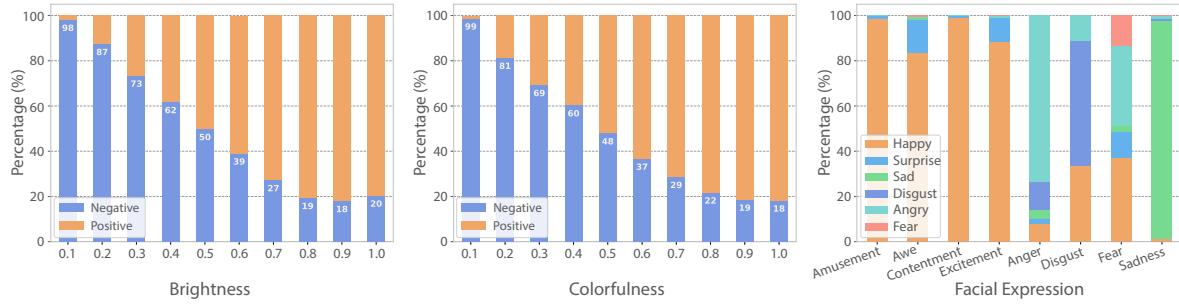


Figure 4: Histogram of brightness, colorfulness and facial expression, where different colors suggest different categories.

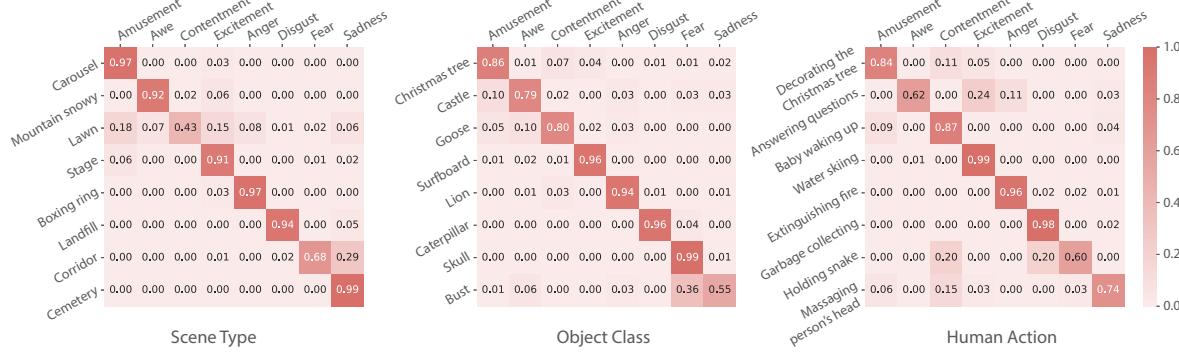


Figure 5: Correlation matrices of scene type, object class and human action, where numbers on the diagonal indicate the relationships between emotions and their top-1 attribute values.

down of each brightness and colorfulness level into negative (blue) and positive (orange) emotions. Indeed, it may be seen that the proportion of images with a positive emotion label increases from left to right.

Our facial expression attribute is built upon Ekman model, where happy is positive, surprise is neutral, and other four (*i.e.*, angry, disgust, fear, and sad) are negative. In this experiment, we would like to see how facial expressions influence visual emotions. In Figure 4, we show the breakdown of facial expressions for different visual emotions. Unsurprisingly, all of the positive emotions exhibit a high correlation with a happy facial expression, while anger, disgust, and sadness are highly correlated with their corresponding facial expressions. Interestingly, for fear, the top facial expression is happy, probably because the image contains a sinister or a spooky smile. The above experiment indicates that people are easily affected by the facial expressions present in the image, a manifestation of *empathy* [8].

Each of the attributes scene type, object class, and human action, may have many different values, as in Figure 3. Obviously, some attribute values are strongly related to emotions (*e.g.*, amusement park, cemetery, laughing, or crying), while others are not, such as sky, plant, tree, or window. To discover emotion-related attribute values, we calculate the co-occurrence between each emotion-attribute pair and adopt the TF-IDF technique [40], where the importance of a value increases when it appears in a specific emotion and

decreases with its appearance in the whole dataset. Figure 5 presents the correlation matrices between each emotion and its top-1 attribute value, where the large number on the diagonal suggests a strong relationship between them, with an average on 0.85 (scene type), 0.86 (object class) and 0.83 (human action). The statistics in Figure 5 are highly consistent with human cognition, indicating that some attribute values are indeed strongly related to emotions. Once a certain attribute value appears, the image is much more likely to evoke the corresponding emotion.

5. Evaluation of EmoSet

5.1. Datasets Comparison

Development in VEA has been greatly limited by the lack of large-scale, high-quality datasets, leading to unsatisfying results in visual emotion recognition, *i.e.*, 59.60% in Emotion6, 70.07% in FI, as shown in Table 3. Consisting of 118,102 carefully annotated images, EmoSet aims at serving as an important dataset in VEA. To verify the quality of EmoSet, we compare it with other datasets, where results are reported in top-1 accuracy (%). We conduct experiments by leveraging both classic convolutional neural networks, *i.e.*, AlexNet [21], VGG-16 [43], ResNet-50 [15] and DenseNet-121 [16], as well as VEA-oriented methods, *i.e.*, WSCNet [53], StyleNet [59], PDANet [61], Stimuli-aware [52] and MDAN [48]. The classic convolutional

Table 3: Comparison on visual emotion recognition of EmoSet versus other VEA datasets with top-1 accuracy (%).

Method	Twitter I-2	Twitter II-2	Flickr-2	Instagram-2	Emotion6-6	FI-8	EmoSet-2	EmoSet-8
AlexNet [21]	75.20	75.63	79.73	77.29	44.19	59.85	89.28	67.80
VGG-16 [43]	78.35	77.31	80.75	78.72	49.75	65.52	93.40	72.27
ResNet-50 [15]	79.53	78.15	82.73	81.45	52.27	67.53	93.48	74.04
DenseNet-121 [16]	80.71	78.99	84.87	83.76	53.79	67.24	92.92	72.32
WSCNet [53]	84.25	81.35	81.36	81.81	58.25	70.07	94.16	76.32
StyleNet [59]	81.50	80.67	85.02	84.53	59.60	68.85	93.93	77.11
PDANet [61]	80.71	77.31	85.36	83.80	59.34	68.05	94.01	76.95
Stimuli-aware [52]	82.28	79.83	85.64	84.90	61.62	72.42	94.58	78.40
MDAN [48]	80.24	83.05	84.26	83.52	61.66	76.41	93.71	75.75

Table 4: Ablation study of attribute module with different backbones on EmoSet, where results are reported in top-1 accuracy (%).

Backbone	w/o pretrained			w/ pretrained		
	w/o attr	w/ attr	Δ	w/o attr	w/ attr	Δ
AlexNet	46.44	55.80	↑ 9.36	67.80	70.09	↑ 2.29
VGG-16	48.51	56.51	↑ 8.00	72.27	74.76	↑ 2.49
ResNet-50	51.48	58.62	↑ 7.14	74.04	76.60	↑ 2.56
DensNet-121	53.09	60.77	↑ 7.68	72.32	74.94	↑ 2.65
Average	49.88	57.93	↑ 8.05	71.61	74.10	↑ 2.50

neural networks are first pretrained on ImageNet [6], then fine-tuned and tested on each dataset respectively, while the VEA-oriented methods are trained and tested following their specific settings. For EmoSet, we split the data into 80% training set, 5% validation set and 15% test set, in accordance with that of FI. Our network is trained by the adaptive optimizer Adam [19], where the learning rate is set as the default one in each method. Our experiments are implemented based on PyTorch [33] and performed on an NVIDIA GTX 3090 GPU. Notably, some datasets are labeled with two (*e.g.*, Twitter I-2) or six (*e.g.*, Emotion6-6) emotion categories, making it unfair to compare them with the eight-category EmoSet. Therefore, we degenerate eight emotions to two sentiments, denoted as EmoSet-2, where amusement, awe, contentment and excitement are *positive* and anger, disgust, fear and sadness are *negative*. In Table 3, EmoSet reaches the best performance in both 2-sentiment and 8-category recognition tasks, compared with other VEA datasets.

5.2. Attribute-aware Visual Emotion Recognition

We propose an attribute module to facilitate visual emotion recognition, as shown in Figure 6, which can be easily attached to a backbone network. Attribute module is built with three branches, namely low-level, mid-level and high-level, to extract different visual information from a given image. The extracted features are then sent to several lightweight convolutional layers, yielding attribute features. Take backbone as main branch, features from other branches are fused with it to jointly predict visual emotions.

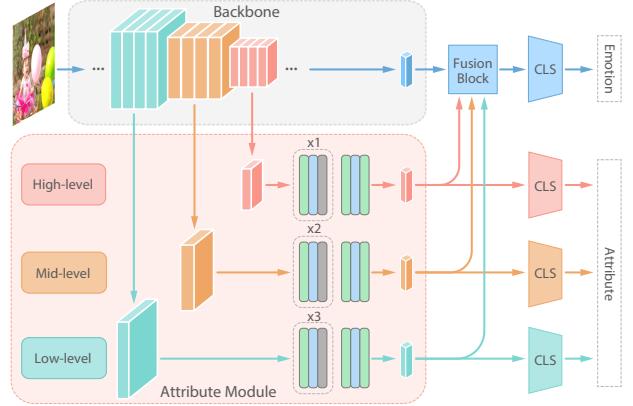


Figure 6: The proposed attribute module.

According to Section 3.2, there are many options for selecting attributes. In our experiments, we choose brightness, scene type and facial expression as representatives, which results are shown in Table 4. Each attribute branch is supervised by its ground-truth label, while the main branch is supervised by the emotion label. The whole network is trained on EmoSet in an end-to-end manner. Table 4 reports the top-1 accuracy (%) under several conditions: with or without pretrained on ImageNet (*i.e.*, w/ pretrained, w/o pretrained), with or without attribute module (*i.e.*, w/ attr, w/o attr) and the differences between them (*i.e.*, Δ). We conduct experiments with several widely-used backbones, including AlexNet, VGG-16, ResNet-50 and DenseNet-121. The results indicate that emotion attribute boosts recognition performance to a large extent, especially when backbone has not been pretrained by ImageNet (*i.e.*, 8.05% on average). The validity of emotion attributes has been verified in the above experiments, which serve as vital auxiliary information to help learn visual emotions.

To verify how attributes assist emotion recognition, we further visualize the attribute features extracted from different branches, *i.e.*, brightness, scene type and facial expression, as shown in Figure 7. Our visualizations are based on test set. Each 2048-dimensional attribute feature is projected to a 2-dimensional vector by using t-Distributed Stochastic Neighbor Embedding (t-SNE) [45], shown as a

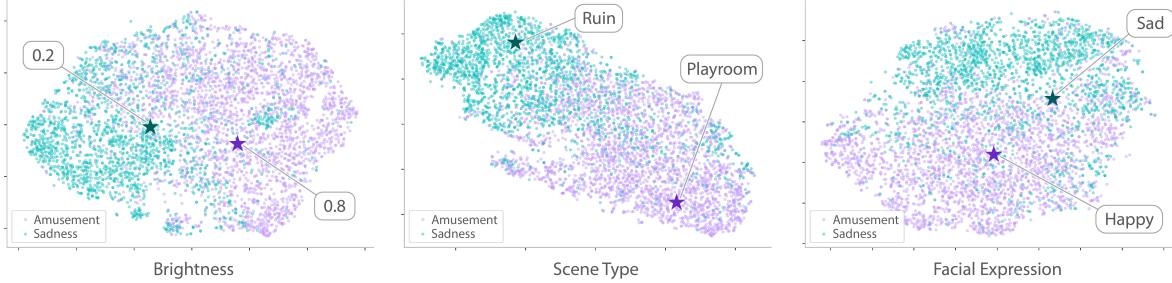


Figure 7: Scatter diagram of brightness, scene type and facial expression, where different colors represent attribute features of different emotions.

Table 5: Cross-dataset generalization between FI and EmoSet, where results are reported in top-1 accuracy (%).

test \ train	w/o pretrained			w/ pretrained		
	FI	EmoSet	Δ	FI	EmoSet	Δ
FI	40.62	35.26	$\downarrow 5.36$	67.53	55.67	$\downarrow 11.86$
EmoSet	24.50	51.48	$\downarrow 26.98$	53.95	74.04	$\downarrow 20.09$
Artphoto	18.23	25.06	-	29.28	32.26	-

data point in the Cartesian coordinate. Taking amusement and sadness as examples, the scatter diagram is designated by two different colors, *i.e.*, purple and green. Obviously, data points can be separated by different emotions, especially in scene type, indicating that attribute features have learned some emotion-related information. We also visualize several emotion-related attribute values by calculating their cluster centers, which is denoted as stars in Figure 7. In scene type, “ruin” and “playroom” are separated with a large distance, locating in sadness and amusement respectively. Different attribute values fall in different emotional areas, which suggests that attribute features have been trained to distinguish both emotions and attributes jointly. Visualization results in Figure 7 further proved the effectiveness of emotion attributes on assisting visual emotion recognition and understanding, which is also consistent with our human cognition.

5.3. Cross-dataset Generalization

To demonstrate the generalization ability of EmoSet, we conduct a cross-dataset validation between EmoSet and the large-scale FI in Table 5. We choose ResNet-50 as our backbone. To validate the performance on a broader sense, we conduct two experiment settings: with or without pretrained on ImageNet, denoted as w/ pretrained, w/o pretrained. Trained on FI, the backbone meets a performance drop of 26.98% (w/o pretrained) and 20.09% (w/ pretrained), compared to the baseline of EmoSet. Conversely, trained on EmoSet, the backbone meets a performance drop of 5.36% and 11.86%, correspondingly. The above results illustrate that EmoSet is more capable to generalize to FI, compare with the opposite. We further conduct validations by introducing a third-party test dataset, Artphoto, which consists

of artistic images. Since Artphoto is a small-scale dataset with 806 images, we only use it for test purpose. In each setting, model trained on EmoSet performs better than that of FI, resulting from the diverse image types in EmoSet, *i.e.*, social and artistic. Consisting of high-quality and diverse images, EmoSet is robust to generalize to other VEA datasets with a good visual emotion representation, which may bring new opportunities to visual emotion recognition.

6. Conclusion

EmoSet is built with three main goals. The first one is to provide a large-scale, diverse and balanced VEA dataset, which may offer new opportunities for visual emotion recognition. Second, we believe that the rich annotated attributes can serve as auxiliary information to boost recognition performance. Most importantly, we hope that the comprehensively designed emotion attributes will encourage VEA researchers to turn their eyes from recognition to understanding, and to dive deep into visual emotion.

The underlying premise of our work is that each image evokes a single type of emotion. In reality, different emotions may be evoked at the same time, considering the subjectivity of human emotions. Besides, our work is built upon Mikels emotion model with eight categories, following previous work. It is obvious that emotions are complex, and it is hard to precisely classify them into only a few discrete types. We will explore more in these directions.

With 3.3 million images tagged with emotions and texts, EmoSet has the potential for weakly supervised learning, vision-language modeling and multi-modal emotion analysis. With rich attribute annotations, EmoSet also holds promise for visual emotion generation and editing.

Acknowledgments: This work was supported in parts by NSFC (62161146005, U21B2023), DEGP Innovation Team (2022KCXTD025), Guangdong Science and Technology Program (2023A1515011440), Shenzhen Science and Technology Program (KQTD20210811090044003, RCJC20200714114435012), Israel Science Foundation (3441/21, 2492/20, 3611/21), and Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ).

References

- [1] Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. Angrybert: Joint learning target and emotion for hate speech detection. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 701–713. Springer, 2021. [2](#)
- [2] Simon Baron-Cohen and Sally Wheelwright. The empathy quotient: an investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders*, 34(2):163–175, 2004. [4](#)
- [3] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232, 2013. [2, 3](#)
- [4] Tobias Brosch, Gilles Pourtois, and David Sander. *The perception and categorisation of emotional stimuli: A review*. Psychology Press, 2010. [2, 4](#)
- [5] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision*, pages 288–301. Springer, 2006. [4](#)
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [5, 7](#)
- [7] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993. [2, 4](#)
- [8] Robert Elliott, Arthur C Bohart, Jeanne C Watson, and Leslie S Greenberg. Empathy. *Psychotherapy*, 48(1):43, 2011. [6](#)
- [9] Florian Eyben, Martin Wöllmer, Tony Poitschke, Björn Schuller, Christoph Blaschke, Berthold Färber, and Nhu Nguyen-Thien. Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car. *Advances in human-computer interaction*, 2010, 2010. [1](#)
- [10] Nico H Frijda. Emotion experience and its varieties. *Emotion Review*, 1(3):264–271, 2009. [4](#)
- [11] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer, 2013. [4](#)
- [12] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. Mscap: Multi-style image captioning with unpaired stylized text. In *CVPR*, pages 4204–4213, 2019. [2](#)
- [13] Alan Hanjalic. Extracting moods from pictures and sounds: Towards truly personalized tv. *IEEE Signal Processing Magazine*, 23(2):90–100, 2006. [2, 3](#)
- [14] Patrick Hanks. Collins dictionary of the english language. 1979. [3](#)
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. [6, 7](#)
- [16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. [6, 7](#)
- [17] Marie Katsurai and Shin’ichi Satoh. Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2837–2841. IEEE, 2016. [3](#)
- [18] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. [4](#)
- [19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [7](#)
- [20] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotic: Emotions in context dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 61–69, 2017. [3](#)
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012. [6, 7](#)
- [22] Pınar Kurt, Kübra Eroğlu, Tubanur Bayram Kuzgun, and Bahar Güntekin. The modulation of delta responses in the interaction of brightness and emotion. *International Journal of Psychophysiology*, 112:1–8, 2017. [2, 4](#)
- [23] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. [4](#)
- [24] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. International affective picture system (iaps): Instruction manual and affective ratings. *The center for research in psychophysiology, University of Florida*, 1999. [2](#)
- [25] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022. [4](#)
- [26] Leida Li, Hancheng Zhu, Sicheng Zhao, Guiguang Ding, and Weisi Lin. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. *IEEE Transactions on Image Processing*, 29:3898–3910, 2020. [2](#)
- [27] Zuhe Li, Yangyu Fan, Bin Jiang, Tao Lei, and Weihua Liu. A survey on sentiment analysis and opinion mining for social multimedia. *Multimedia Tools and Applications*, 78(6):6939–6967, 2019. [2](#)
- [28] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92, 2010. [2, 3](#)
- [29] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior research methods*, 37(4):626–630, 2005. [2, 3](#)

- [30] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 3
- [31] Marvin Minsky. *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon and Schuster, 2007. 1
- [32] Karen Panetta, Chen Gao, and Sos Agaian. No reference color image contrast and quality measures. *IEEE transactions on Consumer Electronics*, 59(3):643–651, 2013. 4
- [33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 7
- [34] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 860–868, 2015. 2, 3
- [35] Rosalind W Picard. *Affective computing*. MIT press, 2000. 1
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 4
- [37] Tianrong Rao, Xiaoxu Li, and Min Xu. Learning multi-level deep representations for image emotion classification. *Neural Processing Letters*, pages 1–19, 2016. 2, 3
- [38] Timothy D Ritchie and Tamzin J Batteson. Perceived changes in ordinary autobiographical events’ affect and visual imagery colorfulness. *Consciousness and cognition*, 22(2):461–470, 2013. 4
- [39] Peter Mark Roget. *Roget’s Thesaurus of English Words and Phrases...* TY Crowell Company, 1911. 3
- [40] Gerard Salton and Clement T Yu. On the construction of effective vocabularies for information retrieval. *Acm Sigplan Notices*, 10(1):48–60, 1973. 6
- [41] Pablo Sánchez-Núñez, Manuel J Cobo, Carlos De Las Heras-Pedrosa, José Ignacio Peláez, and Enrique Herrera-Viedma. Opinion mining, sentiment analysis and emotion understanding in advertising: a bibliometric analysis. *IEEE Access*, 8:134563–134576, 2020. 2
- [42] Abhinav Shukla, Shruti Shriya Gullapuram, Harish Katti, Mohan Kankanhalli, Stefan Winkler, and Ramanathan Subramanian. Recognition of advertisement emotions with application to computational advertising. *IEEE Transactions on Affective Computing*, 2020. 1
- [43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6, 7
- [44] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*, 2018. 4
- [45] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7
- [46] Matthias J Wieser, Elisabeth Klupp, Peter Weyers, Paul Pauli, David Weise, Daniel Zeller, Joseph Classen, and Andreas Mühlberger. Reduced early visual emotion discrimination as an index of diminished emotion processing in parkinson’s disease?—evidence from event-related brain potentials. *Cortex*, 48(9):1207–1217, 2012. 2
- [47] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Large-scale datasets for going deeper in image understanding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1480–1485. IEEE, 2019. 2
- [48] Liwen Xu, Zhengtao Wang, et al. Mdan: Multi-level dependent attention network for visual emotion analysis. In *CVPR*, 2022. 6, 7
- [49] Elaheh Yadegaridehkordi, Nurul Fazmidar Binti Mohd Noor, Mohamad Nizam Bin Ayub, Hannyzzura Binti Affal, and Nornazlita Binti Hussin. Affective computing in education: A systematic review and future research. *Computers & Education*, 142:103649, 2019. 1
- [50] Dingkang Yang, Shuai Huang, et al. Emotion recognition for multiple context awareness. In *ECCV*, 2022. 3
- [51] Jingyuan Yang, Xinbo Gao, Leida Li, Xiumei Wang, and Jinshan Ding. Solver: Scene-object interrelated visual emotion reasoning network. *IEEE Transactions on Image Processing*, 30:8686–8701, 2021. 2, 3
- [52] Jingyuan Yang, Jie Li, Xiumei Wang, Yuxuan Ding, and Xinbo Gao. Stimuli-aware visual emotion analysis. *IEEE Transactions on Image Processing*, 30:7432–7445, 2021. 3, 6, 7
- [53] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *CVPR*, pages 7584–7592, 2018. 2, 3, 6, 7
- [54] Jufeng Yang, Dongyu She, Ming Sun, Ming-Ming Cheng, Paul L Rosin, and Liang Wang. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia*, 20(9):2513–2525, 2018. 3
- [55] Jufeng Yang, Ming Sun, et al. Learning visual sentiment distributions via augmented conditional probability neural network. In *AAAI*, 2017. 3
- [56] Georgios N Yannakakis. Enhancing health care via affective computing. 2018. 1
- [57] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 2015. 3
- [58] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016. 2, 3, 5
- [59] Wei Zhang, Xuanyu He, and Weizhi Lu. Exploring discriminative representations for image emotion recognition with cnns. *IEEE Transactions on Multimedia*, 22(2):515–523, 2019. 6, 7
- [60] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art

- features for image emotion recognition. In *ACMMM*, pages 47–56, 2014. [2](#), [3](#)
- [61] Sicheng Zhao, Zizhou Jia, Hui Chen, Leida Li, Guiguang Ding, and Kurt Keutzer. Pdanet: Polarity-consistent deep attention network for fine-grained visual emotion regression. In *Proceedings of the 27th ACM international conference on multimedia*, pages 192–201, 2019. [6](#), [7](#)
- [62] Sicheng Zhao, Hongxun Yao, Yue Gao, Guiguang Ding, and Tat-Seng Chua. Predicting personalized image emotion perceptions in social networks. *IEEE transactions on affective computing*, 9(4):526–540, 2016. [2](#)
- [63] Sicheng Zhao, Xingxu Yao, et al. Affective image content analysis: Two decades review and new perspectives. *IEEE TPAMI*, 2021. [2](#)
- [64] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. [4](#)
- [65] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018. [4](#)
- [66] Jing Zhu and Paul Thagard. Emotion and action. *Philosophical psychology*, 15(1):19–36, 2002. [4](#)