Exploring emotional representation in deep neural network image emotion classifiers

Konrad Cybulski

March 2019

1 Introduction

For centuries artists have been extremely talented at creating pieces of artwork that convey a range of emotions to those who view them. Extensive research has been conducted into how visual features affect humans emotionally and how these can be used to predict and detect the emotional content of images and text [18, 38]. Due to the subjective and qualitative nature of human emotion, assigning a quantitative measure of emotion to an image is no easy task. Furthermore the ability to computationally recognise the emotional content of an image has wide-ranging applications from classifying posts on social media, to the creation of images, text, and even physical spaces in an emotionally quantifiable way.

Methods for quantifiably representing emotion have been explored thoroughly in the domain of psychology, with continuous multi-dimensional models being used in lieu of a single emotional label . The circumplex model of emotion introduced a two-dimensional space characterised by valence, and arousal; respectively representing positivity or negativity, and the level of excitement associated with it [29]. Such a continuous model is not without flaw, failing to accurately capture more complex emotional states, often those that represent concurrent conflicting sides of a given axis [14]. The complexity of such a continuous representation of emotion has been investigated in depth and extended in various ways such as by Bradley and Lang [3], through the addition of a third dimension: dominance; which is particularly of interest within social dynamics.

The domain of emotion classification has had a particular focus on facial expressions and text [4, 34]. Image emotion classifiers have been explored [13, 18, 5, 6] yet their use has been limited. While humans ability to recognise, label, and discuss the emotive content of an image is not lacking, the capability to computationally classify image emotion, and the underlying patterns learnt by such classifiers is.

As with the use of deep neural network image classifiers such as ResNet, AlexNet, and Inception, the shapes and patterns learnt by such deep learning systems are difficult to extract. Recent work has looked at the use of quality-diverse generative algorithms with such deep classifiers [23, 24]. The underlying patterns learned by the classifier can be surfaced by synthesising images that

maximise various desirable features. This method allowed the exploration of images to which the classifier assigns a given label such as *school bus* or *lighthouse*, enabling a deeper understanding of the visual characteristics inherent to each class.

2 Aims

The aim of this research is to better understand the visual patterns associated with various emotions conveyed in images. This will primarily leverage image emotion recognition architectures explored by Kim et al. [13], in combination with the dataset produced by Zhao et al. [40] containing over 1.4 million images with assigned valence, arousal, and dominance levels derived from their descriptions. Producing an architecture with which valence-arousal (VA) values can be assigned to images forms the basis for this research. Such a platform allows enables the exploration into how VA values are assigned to more complex, multi-faceted and multi-layered images, and the efficacy with which this is done. Furthermore this process, and the patterns learned by it will be better understood through the use of generative processes which maximise given target features in a quality-diverse way [24, 23], or through a generative adversarial approach [31]. In the context of this research, such features include valence, arousal, dominance, happiness, sadness, etc. This will allow a great understanding of the visual patterns that such an architecture learns, and any psychological or artistic parallels that can be drawn. The combination of such a generative system with a classifier of emotion can be further extended to domains such as generative art and text-to-image synthesis, with a focus on emotion-driven image generation.

2.1 Research questions

- How does quantitative emotion representation affect convolutional neural network classifiers of image emotion?
- What class-differentiating visual patterns and characteristics are learned by such classifiers?
- How do such visual characteristics learnt by an image emotion classifier relate to known psychological and artistic understandings of image emotion?

3 Background

3.1 Image emotion recognition

☐ Emotion classification in images: facial expression, general imagery.

- ☐ Methods of classifying emotion in images: single target emotion, discrete categorical likelihood, decomposition into continuous vector (valence-arousal).
- ☐ Discuss paper by Gauthier [11] on conditional GAN and relate to Tan et al. [31] who leverages this technique.

The area of image emotion and sentiment classification has been explored in a number of ways, primarily through image feature analysis derived from art and psychological factors [18]; and more recently using techniques such as deep neural networks [5, 13]. Feature extraction and analysis has been used for various applications such as measuring aesthetic appeal [7, 10, 8] and as an emotional feature vector for sentiment classification [18]. Due to the artistic and psychological underpinnings used by Machajdik and Hanbury [18], the low-level features extracted from images can be understood at a high level. The relationship between an image's emotion and its core artistic components such as balance, harmony, and variety was further explored by Zhao et al. [38], which uses a comparably small feature vector to Machajdik and Hanbury [18], resulting however in a 5% classification increase to state-of-the-art approaches at the time.

Deep neural networks in this domain provide less transparency to the process with which emotions and sentiment are classified compared to feature analysis. The emotional content of an image can be decomposed in various ways. Image databases with singular emotion labels, and adjective-noun pairs (ANP) have been used for the training of deep neural network classifiers [6, 35] with up to 200% performance gains over support vector machine classifiers.

Given the large amount of work conducted into image object classification, producing neural networks such as AlexNet, ResNet, and Inception, transfer learning has been heavily relied upon in various domains of classification and synthesis from images. Techniques used by image emotion recognition have been used extensively in domains such as image sentiment analysis [36], and image-to-text synthesis [32]. The subjectivity and human-dependent nature of emotion has naturally resulted in only small datasets of emotion-labelled images. Through transfer learning, pre-trained image object classifiers can and have been used in the domain of image emotion [13, Wang and Lewis].

Emotion representation Sentiment is typically represented as a binary classification of either positive or negative [35, 6]. Facial and image emotion classification however face a higher complexity problem, where research has typically focused on a subset of potential emotions ranging in size from 7 [1], 8 [18], 11 [Wang and Lewis], and even 19 [20]. Other methods for representing emotion include use of the circumplex model of affective emotion [29, 3]. This model represents emotion as a continuous multi-dimensional space. The original definition by Russell [29] introduces a two-dimensional space defined by valence, and arousal. Valence represents the positive and negative aspect of an emotion; arousal, the excitatory component. In such a model, each emotion label used in common categorical classification resides within the space as seen in 1. The circumplex model has been applied and extended in various ways. The addition

of a third dimension, dominance, introduced by Bradley and Lang [3] allowed the representation of emotion relating to feelings of situational control. Despite the increased freedom of representation in a continuous space, problems exist with the circumplex model [14]. Some of the key issues with such a model is its dichotomous nature, resulting in a comparative failure to capture more complex emotions, particularly emotional states representing conflicting sides of a dimension.

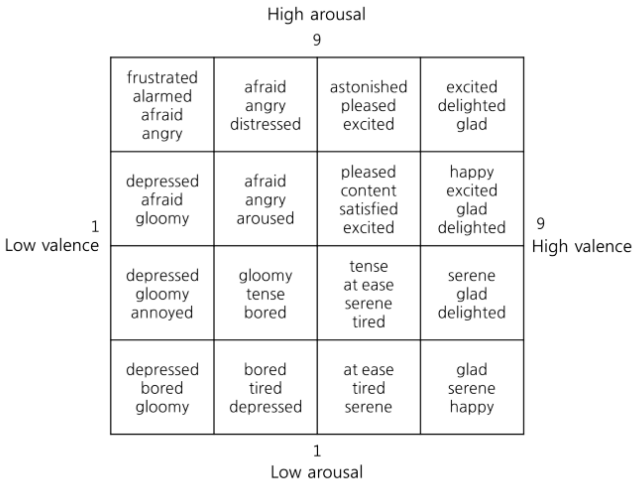

Figure 1: Distribution of emotions associated with levels of valence and arousal determined by DNN classifier produced by Kim et al. [13]

The use of continuous emotion representations, particularly relating to the circumplex model have been explored in image emotion recognition tasks [13, 40, 39]. Regression models produced to predict the valence-arousal (VA) values of given images have shown high accuracy on various datasets. In leveraging pre-trained image classification networks through transfer learning, even smaller datasets (10,000 images) can have high accuracy classification results [13].

3.2 Computational image synthesis

Generative systems have been explored in various domains with numerous techniques. Examples range from the generation of images and art using evolutionary methods [30, 17], to the synthesis of text to describe a given image through the use of deep neural networks [19]. Some of the first *human-in-the-loop* image synthesis systems such as *NEvAr* [17] produced greatly impressive images through evolutionary techniques. Evolutionary art leveraged methods introduced and exemplified by Sims [30] such as those shown in Figure 2. Sims [30] proposed using *Lisp* expressions for genotype definitions, which accepted

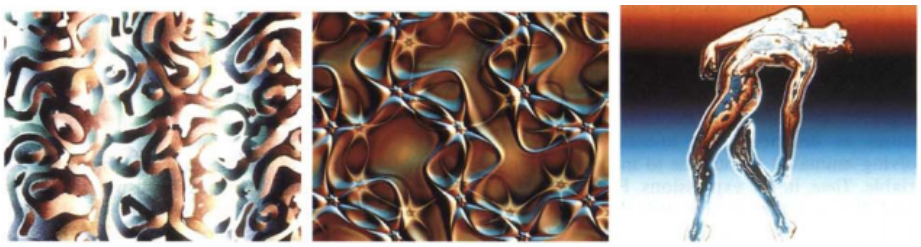

Figure 2: Images generated through the process of interactive evolution introduced by Sims [30]

a coordinate (x, y) which could be evaluated into a grayscale or RGB value producing images. This genotype expression has been used in numerous further research into the process of both supervised and unsupervised image synthesis through evolutionary techniques [17, 30, 8, 9, 28].

Despite the slow nature of the interactive process, Sims [30] and Machado and Cardoso [17] were able to produce images with visually striking characteristics. Ross et al. [28] investigated measures of aesthetics for fitness evaluation in artificially evolving images. This research primarily used observations by Ralph [26], that the distribution of colour gradients in fine art tend towards normal. While the images produced through this method did not meet the level of intricacy and detail as the results of Sims [30] or Machado and Cardoso [17], it represented a self-contained system able to generate appealing imagery without human interaction.

Introduction of the generative adversarial network architecture (GAN) by Goodfellow et al. [12] allowed the process of image generation to depend only on collecting a sufficiently large dataset. Common GAN application has involved the generation of realistic images, such as has been done by Bao et al. [2], where images have been synthesised to fine-detailed target labels such as bird species' and actors. Zhang et al. [37] and Reed et al. [27] have recently explored text to image synthesis, in which detailed descriptions of birds and flowers have been converted into photo-realistic images using the GAN model. Such an architecture has been applied to the area of art synthesis by Tan et al. [31] in which images were generated according to a target genre and artist. Learning from a dataset of countless artworks under various categories, styles, and artists, it was hugely successful in generating images that were stylistically similar to existing art of the target artist/genre. Due to the competitive relationship of the generator and discriminator networks, the patterns learned by the discriminator propagate through the generator network. The discriminator network of the GAN architecture aims to learn patterns and styles from the dataset on which it is trained in order to discriminate between the generated and existing images. As a result, the images generated tend to resemble closely those in the training dataset, which represents a benefit in successfully producing images closely matching the target domain, and a detriment with regard to

the networks potential creativity.

Recent work by Nguyen et al. [24] and Nguyen et al. [23] investigated the use of quality-diverse algorithms for image generation particularly to better understand the patterns learned by deep neural network image classifiers. Quality-diverse (QD) evolutionary algorithms such as Multi-dimensional Archive of Phenotypic Elites (MAP-Elites) [21] and Novelty Search [15, 16] have been developed to address the need for a high quality, yet diverse solution space in related optimisation domains. The use of such QD algorithms has shown great promise in its efficiency and accuracy on a number of hard optimisation problems [25] such as maze navigation [16]. Nguyen et al. [23] and Nguyen et al. [24] use MAP-Elites in conjunction with a pre-trained deep neural network (DNN) image classifier; assigning individual image fitness according to the accuracy with which it is classified. Using the MAP-Elites framework in this context, each dimension of the feature-space represents a classification label, and as such the generated images allow the exploration of label representative patterns and shapes learnt by the classifier. Nguyen et al. [23] leverages such an architecture to show the shallowness with which an image classifier recognises images. Assigning the label of *school bus* to alternating yellow and black lines is a prime example of the way such a network has learnt to differentiate one class from the others. Thus enabling exploration into the inner workings of the DNN classifier by uncovering features that maximise the separation of one label to another. In contrast, Nguyen et al. [24] uses the same architecture to explore the novelty-driven evolutionary path taken by generated images and the potential for such a system in the field of content synthesis. While the conclusions derived from Nguyen et al. [24] and Nguyen et al. [23] contrast greatly, the quality-diverse generative method used to understand the visual components learnt by the classifier show such an architecture's exploratory abilities.

The application of generative systems in the domain of affective computing, particularly with regard to emotion and content synthesis, is limited. *Emotion transfer* was explored by Ali and Ali [1] which involved the transformation of an image's colour and style with the aim of altering its conveyed emotion.

4 Methodology

4.1 Datasets

There exist a few datasets in which non-facial images have labels of their affective emotion on the viewer. Produced recently by [40] is a compilation of 1.4 million images from *Flickr*, each assigned with a respective valence, arousal, and dominance (VAD) according to the circumplex model of emotion [29, 3]. The assigned values are derived from the analysis of each images textual description according to the methods described by Warriner et al. [34]. This dataset also contains the assigned categorical emotion according to the VAD values, and comments written by viewers for each image along with their derived VAD values. This dataset has been chosen due to its large volume of data, and the

method in which each VAD value is assigned. As detailed in Zhao et al. [40], the method for assigning values to each image involved the computational analysis of an image's description, in addition to a human filtering step to ensure the relative accuracy and validity of the assigned values.

As previously discussed, there exist issues with the circumplex model of emotion. There exist shortcomings particularly in both its ability to capture more complex emotion, and the ease with which humans are able to convert felt emotions to a continuous space. The value in using such a large dataset is to overcome any potential error or inaccuracy associated with the assigned VAD values.

4.2 Emotion representation and prediction

Method for representing emotion in this research have been limited to three options based on the dataset: a single categorical label (happy, sad, etc.); a two-dimensional circumplex model (valence-arousal); and a three-dimensional circumplex model (valence-arousal-dominance). A model commonly used for representing emotion in classification tasks is that of a single categorical label, due its simplicity. While this model represents a reductionist view of emotion, omitting a great deal of complexity, it will be the emotional representation used in this investigation. Investigating the applicability of this simplistic model enables comparison between categorical and continuous representations of emotion explored in this work. The more complex models investigated in this research include both the two and three dimensional circumplex models: valence-arousal, and valence-arousal-dominance respectively. Each of these three models will be used as the classification/regression target of a deep neural network (DNN) model.

The DNN model architectures tested will be based on commonly used techniques such as those explored by Kim et al. [13], Chen et al. [5]. This technique, known as transfer learning, involves the repurposing of pre-trained DNN image classifiers by replacing or feeding through the final layers to another network with the desired output shape and proceeding with further training in the given problem domain. Initial investigation into the classification of image emotion will involve the use of such pre-trained DNN classifiers as ResNet, AlexNet, Inception, and VGGNet. Further feature extraction and ensemble methods may be explored, particularly those investigated by Kim et al. [13], Chen et al. [5], and Chen et al. [5].

4.3 Generative architecture for model exploration

In order to explore and better understand the underlying patterns and shapes learnt by the predictor network, a generative architecture is required which allows the generation of images throughout a desired feature-space. In a categorical representation model, the associated feature-space is defined as the coordinate space spanned by the softmax output of the classifier. And that

of the multi-dimensional regression model is spanned by the valence-arousal-dominance coordinate space. The process by which features in the categorical emotion model will be explored involves methods introduced by Nguyen et al. [23, 24]. This process uses the previously discussed quality-diverse MAP-Elites algorithm. Generating images according to the fitness function defined by the accuracy with which it is classified, and storing individuals in the feature-space matrix. This process will result in a multi-dimensional space in which each dimension represents a given emotion categorisation, divided into N bins each with an image that aims to satisfy the emotions with which it was classified.

In the multi-dimensional regression variant, where the given image's VA/VAD values are computed, categorical classification accuracy cannot be used as an objective function. As a result the associated feature-space will be explored using a conditional generative approach similar to that of Tan et al. [31] and Gauthier [11]. With respect to generative adversarial networks, this technique involves the input of a condition vector (target VA/VAD) to both generator and discriminator networks alongside their standard inputs. In doing so a generator can be trained that aims to minimise error between the target condition, and that predicted by the pre-trained regression network. This conditional generative framework can also be used in the categorical representation.

While techniques used by Nguyen et al. [23] and Nguyen et al. [24] to explore DNN image classifiers involved the MAP-Elites algorithm, the task of generating images that maximise an objective function given an input condition can be completed with a conditional GAN approach. The generative system produced as part of this research aims to generate images within a feature-space, that minimise the classification/regression error in order to better understand some key visual patterns and characteristics learned by the predictive network. Since generative systems trained on given sets of data tend toward generating images similar to its training data, this generative exploration aims to incorporate no prior learning into the process. However there exists the possibility of investigating the use of prior generative learning such as that used by Nguyen et al. [22].

4.4 Potential difficulties

5 Expected Outcomes & Contributions

The initial outcome of this project is an image emotion classification system. Multiple types of emotional representation are explored with reference to the classifier, both the categorical and continuous multi-dimensional representation of valence, arousal, and dominance. Comparisons between the accuracy and efficacy of both representations with respect to the classification domain will be explored.

The primary outcome of this research is the creation of a generative system with which the visual patterns and characteristics learned by such a deep image classifier can be explored. Such a generative system will explore various methods

as discussed in *Methodology*, and produce comparisons between them. The content produced by the generative system in order to better understand the visual components learnt by the classifier will be qualitatively analysed in both a technical, artistic and psychological context.

References

[1] Ali, A. R. and Ali, M. (2017). Emotional filters: Automatic image transformation for inducing affect. *arXiv preprint arXiv:1707.08148*.

[2] Bao, J., Chen, D., Wen, F., Li, H., and Hua, G. (2017). Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2745–2754.

[3] Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

[4] Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

[5] Chen, M., Zhang, L., and Allebach, J. P. (2015). Learning deep features for image emotion classification. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 4491–4495. IEEE.

[6] Chen, T., Borth, D., Darrell, T., and Chang, S.-F. (2014). Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*.

[7] den Heijer, E. and Eiben, A. (2010a). Using aesthetic measures to evolve art. In *IEEE Congress on Evolutionary Computation*, pages 1–8. IEEE.

[8] den Heijer, E. and Eiben, A. (2011). Evolving art using multiple aesthetic measures. In *European Conference on the Applications of Evolutionary Computation*, pages 234–243. Springer.

[9] den Heijer, E. and Eiben, A. (2013). Maintaining population diversity in evolutionary art using structured populations. In *2013 IEEE Congress on Evolutionary Computation*, pages 529–536. IEEE.

[10] den Heijer, E. and Eiben, A. E. (2010b). Comparing aesthetic measures for evolutionary art. In *European Conference on the Applications of Evolutionary Computation*, pages 311–320. Springer.

[11] Gauthier, J. (2014). Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2.

[12] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[13] Kim, H.-R., Kim, Y.-S., Kim, S. J., and Lee, I.-K. (2018). Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Transactions on Multimedia*, 20(11):2980–2992.

[14] Larsen, R. J. and Diener, E. (1992). Promises and problems with the circumplex model of emotion.

[15] Lehman, J. and Stanley, K. O. (2008). Exploiting open-endedness to solve problems through the search for novelty. In *ALIFE*, pages 329–336.

[16] Lehman, J. and Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223.

[17] Machado, P. and Cardoso, A. (2000). Nevar–the assessment of an evolutionary art tool. In *Proceedings of the AISB00 Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science, Birmingham, UK*, volume 456.

[18] Machajdik, J. and Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92. ACM.

[19] Mathews, A. P., Xie, L., and He, X. (2016). Senticap: Generating image descriptions with sentiments. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[20] Mohammad, S. and Kiritchenko, S. (2018). Wikiart emotions: An annotated dataset of emotions evoked by art. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

[21] Mouret, J.-B. and Clune, J. (2015). Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*.

[22] Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, pages 3387–3395.

[23] Nguyen, A., Yosinski, J., and Clune, J. (2015a). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

[24] Nguyen, A. M., Yosinski, J., and Clune, J. (2015b). Innovation engines: Automated creativity and improved stochastic optimization via deep learning. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, pages 959–966. ACM.

[25] Pugh, J. K., Soros, L. B., and Stanley, K. O. (2016). Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40.

[26] Ralph, W. (2006). Painting the bell curve: The occurrence of the normal distribution in fine art.

[27] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

[28] Ross, B. J., Ralph, W., and Zong, H. (2006). Evolutionary image synthesis using a model of aesthetics. In *2006 IEEE International Conference on Evolutionary Computation*, pages 1087–1094. IEEE.

[29] Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

[30] Sims, K. (1993). Interactive evolution of equations for procedural models. *The Visual Computer*, 9(8):466–476.

[31] Tan, W. R., Chan, C. S., Aguirre, H. E., and Tanaka, K. (2017). Artgan: Artwork synthesis with conditional categorical gans. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3760–3764. IEEE.

[32] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

[Wang and Lewis] Wang, Y. and Lewis, M. Arttalk: Labeling images with thematic and emotional content.

[34] Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

[35] Yang, J., She, D., Sun, M., Cheng, M.-M., Rosin, P. L., and Wang, L. (2018). Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia*, 20(9):2513–2525.

[36] You, Q., Luo, J., Jin, H., and Yang, J. (2015). Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[37] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915.

[38] Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.-S., and Sun, X. (2014). Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 47–56. ACM.

[39] Zhao, S., Yao, H., Gao, Y., Ji, R., and Ding, G. (2017). Continuous probability distribution prediction of image emotions via multitask shared sparse regression. *IEEE Transactions on Multimedia*, 19(3):632–645.

[40] Zhao, S., Yao, H., Gao, Y., Ji, R., Xie, W., Jiang, X., and Chua, T.-S. (2016). Predicting personalized emotion perceptions of social images. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1385–1394. ACM.