Literature Review: Synthesising emotion-driven images

Konrad Cybulski

May 2019

Contents

1 Aims and Scope

The primary focus of the proposed research is to produce an image emotion classifier, and to further leverage it in the process of emotion-driven image synthesis. The two core themes explored are image generation techniques, and processes for classifying and generating emotional content. As a result this literature review aims to explore and understand the progression of knowledge in the fields of emotion representation and classification, computational image synthesis, and affective content synthesis. All of which represent the multiple facets of the proposed body of research with which an investigation into the use of image emotion classifiers for emotion-driven image generation will be explored.

2 Emotion representation

The field of emotion classification has surged in recent years given a popularity and rising interest in facial emotion recognition. Other research has focused on exploring ways in which emotion can be recognised in images, text, and more abstract content. Underlying these two areas of emotion recognition is the methods for quantitatively representing emotion in both a meaningful, and accurate way. This section will discuss the evolution of these aspects of emotion classification, in addition to their commonalities, and differences.

The computational recognition of emotion has been explored in countless studies and projects, in both the context of image emotion recognition [24, 48, 16] and facial emotion classification [29]. However a core component and key difference between a large number of such bodies of research is the method by which emotion is represented. Two of the most common representations represent a discrete, and continuous approach.

2.1 Categorical

Discrete emotion representations generally involve the categorisation of emotion to a series of labels. Discrete approaches represent the method used in a large number of papers [24, 1, 43, 28] however the number of emotion labels explored varies greatly. An example of the size of emotion label subsets used in studies of image emotion classification using such a discrete model are 7 [1], 8 [24], 11 [43], and even 19 [28]. Due to this lack of consistency in emotion classification targets, such studies often repeat similar data gathering and classifier training methods again for their own use. Despite the consequences of inconsistency in the emotion labels chosen, due to the simplicity of discrete categorical emotion assignment, data gathering can be performed with ease compared to continuous methods of emotion representation. In order to create large image datasets of labelled images can be completed with greater ease when the number of such labels is reduced. However there remains difficulty in labelling images with respect to emotion given the inherent variation in the emotion felt by someone

when both viewing an abstract image, or accurately determining the emotion expressed by a facial expression.

The categorical representation is the simplest method for emotion classification, due not only to the consequent ease with which data can be gathered, but due to the extensive research surrounding methods for computational classification of content. Many techniques for prediction, both for regression and classification rely on large datasets, something more easily available and more accurate for categorical data. Research in psychology has understood the increased error associated with continuous measurement tasked performed by humans in comparison to categorical classification [14]. With the human brain's attempt to sort perceived objects and situations into learned categories, it has been shown to warp continuous variables and scales to do so. As a result there exists an increased error in human measurement of continuous, compared to categorical variables.

2.2 Continuous dimensional

The aforementioned difficulty associated with labelling images according to their respective emotional content is accentuated with the introduction of a dimensionally continuous representation of emotion. While proposed representations vary, the most recognised basis for many continuous models is the circumplex model of emotion [39]. The circumplex model of emotion introduced by Russell [39] asserts that emotion can be measured in terms of two continuous variables: valence, and arousal. Valence measures the positivity or negativity associated with an emotion; and arousal measures the excitement associated with it. For example, the discrete emotion label of **happiness** could be represented in the continuous valence-arousal (VA) space with high-valence, medium-arousal; while the label **depressed** would translate to low-valence, low-arousal; and **relaxed** being medium-valence, low-arousal. While this continuous dimensional model allows for greater continuity in measuring emotions, it is not without fault [19]. The primary limitations of such a model involve it's likelihood of misinterpretation particularly when considering the labels of each dimension, and any information loss arising from reducing more complex emotions to a two-dimensional space. One way in which the circumplex model of affect introduced by Russell [39] has been extended to address its weaknesses is through the introduction of a third-dimension: Dominance, the amount of control associated with an emotion. Despite its limitations, the circumplex model has been used extensively in the field of psychology [3, 44], and in computational emotion classification to a limited extent [50].

While both models for the representation of emotion have trade-offs, both suffer from issues related to cross-cultural differences in emotion expression and recognition. Cultures inherently differ with respect to how emotions are both felt and expressed [25]; this is becomes increasingly evident when attempting to classify the affective emotion embodied and expressed by more abstract content such as imagery and sound. The popular image-emotion dataset used for classification known as the International Affective Picture System (IAPS) was shown

to have a significantly different valence-arousal assignment for up to 31.74% of images between Chinese and American young adults [15]. Valence-arousal values assigned to images often vary when considering a group of people. To address this potential for error in reporting the subjective emotion imparted by content, the self-assessment manikin (SAM) was introduced [18]. SAM, as can be seen in Figure 1, was developed to aid in the evaluation of emotion, particularly its translation into the commonly used three dimensions of the circumplex model of affect: valence, arousal, dominance. It has proved itself as more accurate and effective than other methods of emotion self-assessment while being less complex [3].

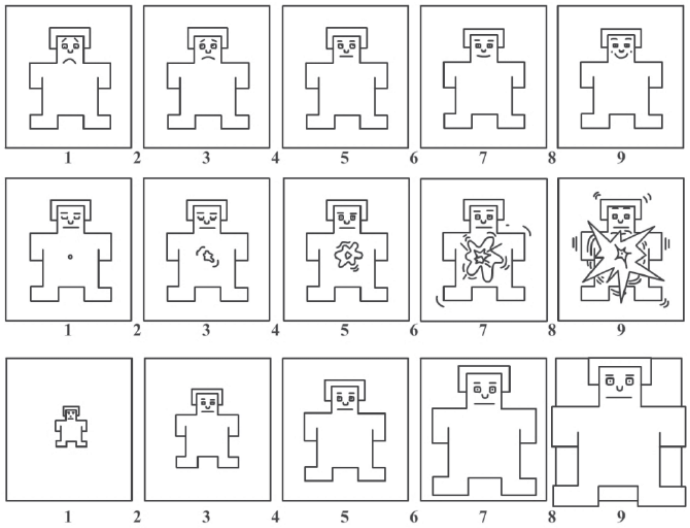

Figure 1: The self-assessment manikin: a guide for reporting individual emotional affect in the three dimensions valence (top), arousal (middle), and dominance (bottom) as introduced by Lang [18]

While both discrete and continuous methods of representing emotion have their respective benefits and trade-offs, the choice to use one over another often depends on the way in which the data gathering process for content emotion classification is performed. These two representations of emotion can be translated between, since all categorical emotions can be converted to the continuous circumplex model, and the inverse. However due to the higher complexity nature of the continuous model, even with the use of SAM, dimensional representations are more difficult to gather on a large scale in comparison to categorical emotions.

3 Image classification

Emotion classification has been researched in depth particularly with respect to two sub-domains: facial emotion recognition, and content emotion classification (image, text, sound). Facial emotion recognition has been more heavily researched in this field when compared to content emotion classification. While the applications of content emotion classifications seem to be few in comparison with those of facial recognition, the field has widely been explored in the context of text emotion classification, with image and sound emotion classification trailing behind.

Emotion classification has itself been a domain built upon the foundation of image classification techniques. From hand-crafted feature decomposition, to neural network architectures, the emotional content of images and text rely heavily on findings in other domains such as image object classification and text sentiment analysis.

3.1 Image object classification

As the basis for many further computer vision applications, image object classification has been a primary focus of research in the field of machine learning. The techniques associated with such tasks have for the most part in recent years involved the use of deep convolutional neural networks, which have proven to be state-of-the-art classifiers, with several high performing architectures such as ResNet, AlexNet, and Inception being used as staple image classifiers in other domains [33]. While neural networks and their use date back to the 1980s, it was only as computing power sufficiently increased in the 21st century that neural network architectures could be viably trained, and used [36].

Up until such time as neural network architectures could be fully leveraged, the majority of image classification tasks lent on manually deriving image features for use in other more traditional classification models (regression, support vector machines, etc.). As a result the limiting factor of such classifiers were predictors derived from extracted image features. The use of such extracted image features has its benefits however. Their primary advantage lies particularly in relation to such features being more interpretable and human-understandable. The features extracted from neural network classifiers however are often abstract and may lack any human-comprehensible parallels despite their comparative predictive advantage. While it is often seen to be a black-box, the features learnt by a deep convolutional neural network, or any neural network architecture for that matter, do represent a repeatable and content-dependent image feature representation. As such the ability to reuse such neural networks and their feature extraction capabilities has been thoroughly explored in various bodies of work [16, 33, 46]. This practice is known as transfer learning.

3.2 Neural networks

This section will give a brief overview of neural networks and their application in image and content classification. The principles underlying neural networks are simple, yet their ability to capture and learn highly complex non-linear representations of information has resulted in their popularity. The fundamental building block of neural networks is the neuron, which itself is a parameterised mathematical function with which a range of inputs can be mapped to a single, or multiple outputs. Examples of such mappings range from simple mathematical functions such linear or logistic regression, to more complex applications involving convolution and functions of input sequences.

Convolutional neural networks represent current state of the art when using images or text for classification or prediction. First introduced by Krizhevsky et al. [17] as part of the ImageNet image classification competition, it exceeded the state-of-the-art at the time in classification accuracy. The way in which such an architecture processes image input data is through convolutional layers. As depicted in figure 2, each neuron in a convolutional layer takes as input, a spatial subset of the layer before it. The result is layers of neurons that have a greater deal of local spatial awareness when compared to the most commonly used fully connected layers. The spatial constraints associated with these layers lends itself to the domain of image processing particularly given the highly spatial relationship of image pixels.

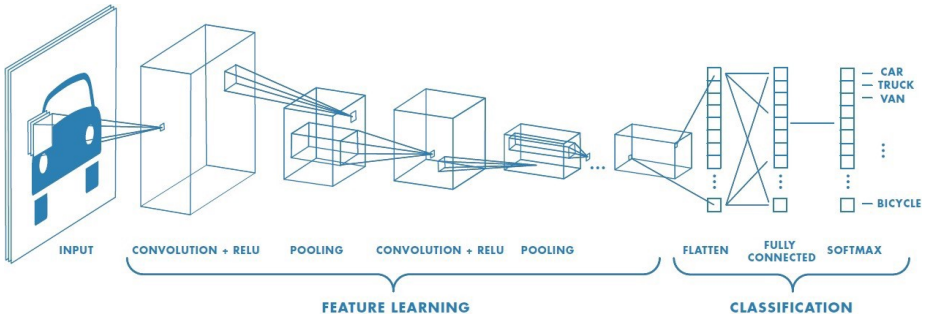

Figure 2: Visual representation of most convolutional neural network image object classifier structures as introduced by Krizhevsky et al. [17]

3.3 Transfer learning

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 of images [33]. Each of these commonly used object classification architectures have been originally trained on common baseline image object classification datasets such as CIFAR-10, CIFAR-100, and ImageNet. These are three of the most common and widely-used image object classification datasets,

in which images fall into one of 10, 100, and 1000 categories respectively. As a result, the number of outputs of neural networks trained on these datasets is the total number of predictable categories. Each of these pre-trained image classifiers can be used in other domains of machine learning by omitting the last layer of their architectures, which represent the object classification layer, and feed the previous feature layer into a new architecture. In doing so, the information held within the neural network can be repurposed, and the learnings reused in other domains of application.

The use of transfer learning in numerous domains has increased the general ability to produce highly accurate networks even with comparatively small datasets.

Not only do the feature layer dimensions of each architecture differs, the features they themselves represent vary. As a result, using stacked predictor architectures, in which multiple feature vectors from multiple neural network classifiers are used as input to a further predictor network [16].

Techniques used by image emotion recognition have been used extensively in domains such as image sentiment analysis [46], and image-to-text synthesis [42]. 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 [16, 43].

3.4 Image emotion classification

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 [24]; and more recently using techniques such as deep neural networks [4, 16]. Feature extraction and analysis has been used for various applications such as measuring aesthetic appeal [6, 10, 7] and as an emotional feature vector for sentiment classification [24]. Due to the artistic and psychological underpinnings used by Machajdik and Hanbury [24], 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. [48], which uses a comparably small feature vector to Machajdik and Hanbury [24], 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 [5, 45] with up to 200% performance gains over support vector machine classifiers. The artificial neural network approach to general image classification domains has exploded in popularity when compared to feature decomposition approaches used by Machajdik and Hanbury [24]. This is largely due to their learning ability, particularly in recognising "hierarchical representations" [22] of image features;

which provides greatly improved performance when compared with manually crafted metrics derived through image decomposition. This hands-off approach to image feature vector decomposition has further lent itself to techniques such as transfer learning.

The use of continuous emotion representations, particularly relating to the circumplex model have been explored in image emotion recognition tasks [16, 50, 49]. 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 [16]. Recent datasets produced for image emotion recognition have used the valence-arousal-dominance model due to its continuity [50]. While categorical classification is more easily verified by humans, training predictive models with data that has an element of noise and uncertainty benefits from both continuity and volume. This is a key advantage of the circumplex model as applied by Kim et al. [16] and Zhao et al. [50] in image emotion recognition over categorical classification.

4 Computational image synthesis

Computer-generated images (CGI), has been a widely used in the film industry, as well as in countless other domains such as gaming, simulation, and art. CGI has, for the most part, involved human interaction, and human-controlled image and model generation. However generative systems and methods for both supervised, and unsupervised image synthesis have evolved in recent years with the increased use of evolutionary algorithms, and in more recent times, neural networks. While extensive research has been conducted into the generation of visually aesthetic images, applications of image synthesis extend to the synthesis of text to describe a given image [26], as well as the generation of an image according to a target caption [37, 47].

4.1 Evolutionary computing

Some of the first methods for image generation focused on the synthesis of visually appealing images. While often using human-in-the-loop systems, visually striking and aesthetic images were the goal of methods introduced by Sims [40] and Machado and Cardoso [23] involving evolutionary techniques. *NEvAr* [23] was one of the first such image synthesis systems able to produce greatly impressive images through evolutionary techniques. Evolutionary art leveraged methods introduced and exemplified by Sims [40], producing images such as those shown in Figure 3. Sims [40] proposed using *Lisp* expressions for genotype definitions, which map a coordinate (x, y) into a grayscale or RGB value. This genotype representation leveraged extensive research done into the use of evolutionary computing for optimisation problems. This genotype expression has been used in numerous further research of both supervised and unsupervised

image synthesis through evolutionary techniques [23, 40, 7, 8, 38]. However the way in which NEvAr and Sims evolved images involved the manual process of selecting individuals in the population they deemed to be of higher fitness than the rest.

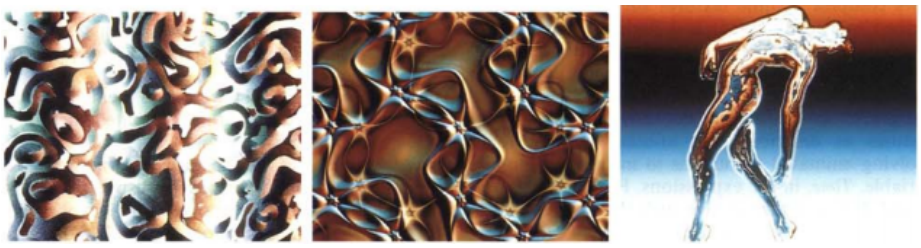

Figure 3: Images generated through the process of interactive evolution introduced by Sims [40]

Evolutionary computing techniques for image synthesis have been able to produce increasingly interesting and appealing imagery and artwork. While the genotype representation introduced by Sims [40] was further used by countless bodies of research, there exist countless other unique methods for image synthesis. Line-drawing as exemplified by McCormack and Bown [27] has built upon evolutionary techniques, using a collection of individuals interacting in real-time to synthesis artwork. In traditional well-mixed [40] or distributed population [8] each individual represents an image, or a model for generating one; however the line-drawing model of image synthesis acts like a swarm of individuals, each drawing on a canvas as they move through the space. This real-time evolutionary image drawing mechanism, while producing visually interesting images, has not been explored with regard to its use in other application domains.

4.2 Measures of aesthetics

Despite the slow nature of the interactive process, Sims [40] and Machado and Cardoso [23] were able to produce images with visually striking characteristics. Ross et al. [38] investigated measures of aesthetics for fitness evaluation in artificially evolving images. This research primarily used observations by Ralph [35], that the distribution of colour gradients in fine art tend toward normal. While the images produced through this method did not meet the level of intricacy and detail as the results of Sims [40] or Machado and Cardoso [23], it represented a self-contained system able to generate appealing imagery without human interaction.

Measures of aesthetics have been explored and multiple have been derived using information about fine art [35], and others measuring levels of symmetry, and even complexity measures according to image compression ratios [6]. Work by den Heijer and Eiben [9] performed a comparison of seven measures of aesthetics, comparing even some of the most popular metrics such as the

Ralph bell curve [35]. The primary finding of this work showed that the visual styles and nature of the images depended heavily on the given aesthetic metric used to determine fitness. Particularly given that the process by which images are generated is an optimisation problem, maximising the aesthetic value as measured by the given metric, the metric used has an immense impact on the resulting styles. However it was found that various pairs of metrics were correlated, resulting in images with highly similar characteristics when using one metric or the other. den Heijer and Eiben [9] also further explored the multi-objective optimisation problem of image synthesis when using combinations of aesthetic measures. The multi-objective optimisation variant of this research showed an increased aesthetic appeal of the images produced particularly with non-correlated metrics.

4.3 Quality-diverse algorithms

Recent work by Nguyen et al. [32] and Nguyen et al. [31] 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) [30] and Novelty Search [20, 21] have been developed to address the need for a high quality, yet diverse solution space in related optimisation domains. The type of problems QD algorithms aim to address include primarily those in which a multitude of solutions exist within a multi-objective space, however the degree to which each objective is desired may vary. Thus algorithms such as MAP-Elites aims to maximise a given fitness function, while maintaining an N-dimensional feature space, where each dimensional represents the feature-specific fitness of a solution. The MAP-Elites algorithm results in not just a single or set of high fitness solutions, but a collection of such high fitness solutions spatially distributed over the desired feature space.

The use of QD algorithms has shown great promise in its efficiency and accuracy on a number of hard optimisation problems [34] such as maze navigation [21]. Nguyen et al. [31] and Nguyen et al. [32] 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. [31] 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. [32] 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. [32] and Nguyen

et al. [31] contrast greatly, the quality-diverse generative method used to understand the visual components learnt by the classifier show such an architecture's exploratory abilities. This technique for understanding the patterns learnt by such a classifier has not been explored in the context of regression.

4.4 Generative adversarial neural networks

Neural networks, while having been applied and researched extensively with regard to prediction and classification, as discussed in Section 3.2, have recently shown exceedingly interesting and even visually realistic results with the generative adversarial architecture. Their use is however limited with regard to image synthesis using more traditional feed-forward network architectures due to the difficulty in converting such networks into higher level image generators. Limited work has been performed using image classifiers in conjunction with evolutionary computing techniques by Nguyen et al. [32] as discussed previously.

The introduction of the generative adversarial network architecture (GAN) by Goodfellow et al. [13] 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, including work by Bao et al. [2] where images have been synthesised to fine-detailed target labels such as bird species' and actors. Zhang et al. [47] and Reed et al. [37] 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. [41] in which images were generated according to a target genre and artist. Learning from a dataset of countless artworks in 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, an advantage when similarity and realism to existing data is desired, and a detriment when generative creativity is a target attribute.

5 Affective content synthesis

The application of generative systems in the domain of affective computing, particularly with regard to emotion and content synthesis, is limited. Sentiment-driven examples of generative systems include image captioning according to target sentiment [26]. The task of describing an image was extended from a traditional GAN approach through the addition of a sentiment target input. The method used to train such a generative system involved the conditional

GAN architecture as described by Gauthier [12]. A similar technique was used in style-driven image captioning (factual, romantic, humorous) in combination with a long short-term memory (LSTM) neural network model Gan et al. [11]. In the context of image-to-image synthesis, *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.

6 Conclusion

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] 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.

[5] 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*.

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

[7] 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.

[8] 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.

[9] den Heijer, E. and Eiben, A. (2014). Investigating aesthetic measures for unsupervised evolutionary art. *Swarm and Evolutionary Computation*, 16:52–68.

[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] Gan, C., Gan, Z., He, X., Gao, J., and Deng, L. (2017). Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146.

[12] 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.

[13] 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.

[14] Harnad, S. (2003). Categorical perception.

[15] Huang, J., Xu, D., Peterson, B. S., Hu, J., Cao, L., Wei, N., Zhang, Y., Xu, W., Xu, Y., and Hu, S. (2015). Affective reactions differ between chinese and american healthy young adults: a cross-cultural study using the international affective picture system. *BMC psychiatry*, 15(1):60.

[16] 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.

[17] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

[18] Lang, P. J. (1980). Behavioral treatment and bio-behavioral assessment: Computer applications.

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

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

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

[22] Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

[23] 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.

[24] 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.

[25] Markus, H. R. and Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological review*, 98(2):224.

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

[27] McCormack, J. and Bown, O. (2009). Life's what you make: Niche construction and evolutionary art. In *Workshops on applications of evolutionary computation*, pages 528–537. Springer.

[28] 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)*.

[29] Mollahosseini, A., Chan, D., and Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE.

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

[31] 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.

[32] 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.

[33] Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

[34] 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.

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

[36] Rawat, W. and Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449.

[37] 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*.

[38] 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.

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

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

[41] 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.

[42] 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.

[43] Wang, Y. and Lewis, M. (2017). Arttalk: Labeling images with thematic and emotional content.

[44] 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.

[45] 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.

[46] 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*.

[47] 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.

[48] 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.

[49] 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.

[50] 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.