Finetuning Convolutional Neural Networks for Visual Aesthetics

Yeqing Wang[1,3]
wyq@ccit.js.cn

Yi Li[2,3,4]
yi.li@tri.global

Fatih Porikli[3,4]
fatih.porikli@anu.edu.au

[1]Changzhou College of Information Technology, China
[2]Toyota Research Institute, United States
[3]National ICT Australia (NICTA), and [4]Australian National University, Australia

***Abstract*—Inferring the aesthetic quality of images is a challenging computer vision task due to its subjective and conceptual nature. Most image aesthetics evaluation approaches focused on designing handcrafted features, and only a few adopted learning of relevant and imperative characteristics in a data-driven manner. In this paper, we propose to attune Convolutional Neural Networks (CNNs) for image aesthetics. Unlike previous deep learning based techniques, we employ pretrained models, namely AlexNet [12] and the 16-layer VGGNet [20], and calibrate them to estimate visual aesthetic quality. This enables exploiting automatically the inherent information from much larger scale and more diversified image datasets. We tested our methods on AVA and CUHKPQ image aesthetics datasets on two different training-testing partitions, and compared the performance using both local and contextual information. Experimental results suggest that our strategy is robust, effective and superior to the state-of-the-art approaches.**

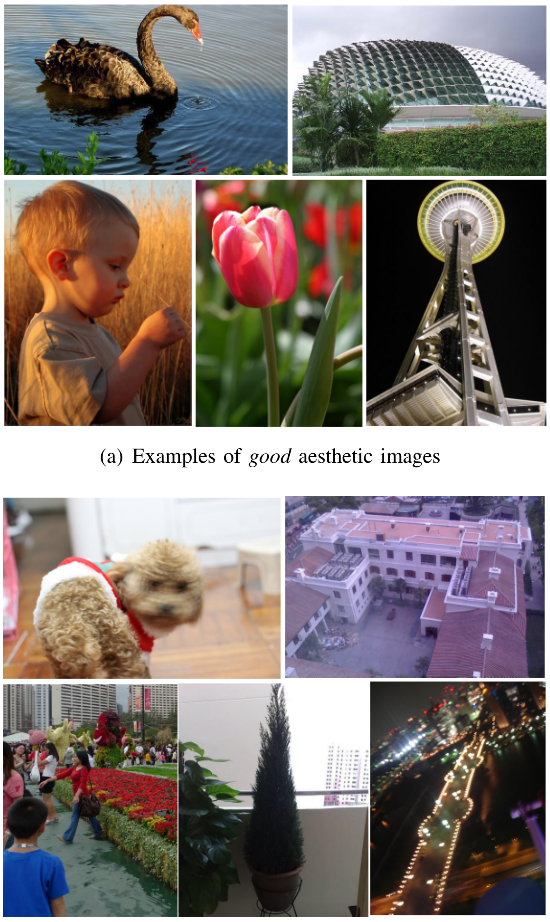
***Index Terms*—Deep learning, visual aesthetics**

I. Introduction

“Beauty is in the eye of the beholder” yet is it possible that the beholder to be a computer? It is evident that we as human beings rely on various visual cues to interpret whether an image is beautiful or not. In order to approximate human perception in this sense, visual aesthetics classification aims to discover such measurable visual cues from a corpus of images and corresponding human responses, and then label given images automatically with binary attributes of good and bad quality.

Figure 1 shows illustrative examples of images with different visual aesthetics. As one can see, the categorization is often subtle and subjective. From a computer vision point of view, this indicates that visual modeling and feature extraction would be intricate and convoluted as much as they are important and essential for discovering measurable cues, which makes image aesthetics analysis an interesting and confounding task.

In recent years, many computational approaches have been proposed. Most of these approaches postulate the image aesthetic estimation task on predefined features [4], [15], [22], [18], [3], [2], [11], [16] that are handcrafted according to holistic image quality attributes and commonly speculated photography rules, such as sharpness, contrast, rule of thirds, visual weight balance, color schemes and so on. However, it is not trivial to designate all elements of measurable visual aesthetics because of the complexity of the attributes and rules. As a result, usability of them becomes severely limited. For instance, [11] applied only portraiture, and [22] used only scenic photographs in their designs.


(a) Examples of *good* aesthetic images

(b) Examples of *bad* aesthetic images

Figure 1. Examples of good and bad quality images from different datasets. As we can see, there are no hand-crafted features that can characterize the difference between these two sets of images easily.

To overcome this shortcoming, generic image descriptors are considered. Numerous generic image descriptors, such as

Bag-of-Words (BOW) [21] and Fisher Vector (FV) [23], have been successfully incorporated for semantic labeling tasks. Several studies demonstrated that these descriptors can learn visual properties efficiently and outperform rule-based features [15]. Nevertheless, these generic features are also handcrafted thus not optimal for visual aesthetic classification task.

In recent years, the emergence of deep learning networks brought image aesthetic assessment to a new era. In [14], [5], visual aesthetics features are learned automatically from images using ordinary neural networks. One major issue with these networks is that they are trained from scratch using comparably small and restricted image aesthetics datasets. While this straightforward training strategy can be regarded as acceptable, it has the limitation of failing to take the advantage of the information available in more diversified datasets. Moreover, the vast variation in visual content strongly requires the network topology to be substantially deep, yet use of small datasets causes serious over-fitting and unstable training convergence behavior of deep networks.

Leveraging on deep learning networks, here we propose utilizing pretrained deep Convolution Neural Networks (CNNs) to overcome the problem of insufficient training data. While borrowing the well-trained models over extensive yet classification oriented datasets, we recalibrate them over small image aesthetics datasets to learn automatically the most indicative features.

To the best of our knowledge, finetuning CNNs for visual aesthetic classification has not been systematically studied. The rich information contained in large and diversified datasets has potential to improve visual aesthetics classification. Motivated by this, [5] also adopts a pretrained network. However, this method uses the pretrained network solely as a feature extractor followed by a binary classifier to classify the images.

Several recent studies suggest that it may be practical to update the network and adapt it to the specific task, which is called as "finetuning". This motivated us to investigate finetuning based solutions for visual aesthetic classification.

We retrain our deep CNNs leveraging on the base of AlexNet model [12] and the VGGNet model [20]. We evaluated our method on two image aesthetics datasets. We first train and test on the popular CUHKPQ dataset. We also test our methods on AVA, a much larger scale aesthetic database [17]. In each dataset, two different cross-validation settings are used. The experiments show that the refined deep VGGNet model provides the state-of-the-art results.

II. Related work

A. Image aesthetics analysis

Quantification of aesthetic quality from images is a formidable task due to the fact that image aesthetics is subjective. Factors not only include the photographic technique but also the opinions of observers. Moreover, the inherent semantic gap between low-level visual features and high-level human-oriented semantics also contributes to this problem [8].

In spite of these challenges, many researchers investigated computationally assessing the aesthetics of images in terms of some common principles and rules-of-thumbs in photography. According to the literature [1], [19], image attributes about colorfulness, composition, spatial organization, depth, etc. can be potentially influential to the response to aesthetics.

Datta et al. [4] considered certain photography rules such as the color tones, saturation, texture, depth, etc. and extracted 56 visual features with respect to these aesthetic rules. Finally, it trained a classifier to distinguish pictures between high and low aesthetic values. Ke et al. [10] designed high level semantic features to measure the perceptual differences between professional photos and low quality snapshots. They reported distinguishing factors between the two types of photos including simplicity, realism, blur, contrast. Based on the perceptual criteria, they manually extracted 5 features to measure the image quality. Aydin et al. [2] focused on rating image aesthetic attributes rather than detecting image distortions. They computed an aesthetic signature from a single image that comprised calibrated ratings of meaningful aesthetic attributes and provided an objective basis for aesthetic evaluation.

Finally, image attributes of sharpness, depth, clarity, tone and colorfulness are chosen to represent the aesthetic signature because they can be expressed algorithmically, and are closely related to photographic principles they claim to model. Marchesotti et al. [15] assessed the aesthetic quality in different ways. They proposed to use generic image features which had been successfully used for semantic tasks to assess the image aesthetics. The generic descriptors like BOW and FV, which encodes the distribution of local statistics, has been shown to capture some aesthetic properties successfully in their experiments.

B. Deep learning for image aesthetics analysis

In recent years, deep learning based methods achieved significant success in computer vision, with image classification being the most outstanding application. Since it was applied on large datasets [12], the deep learning approaches have been used to address key computer vision tasks such as object recognition and object detection. Similar work, e.g. [26], also indicates that deep convolutional neural networks provide improvements in processing not only images and video but also speech and audio.

Finetuning is an optimization strategy of deep learning based methods. For example, [6] suggested that a neural network can form a sufficient generative model of the joint distribution of images and their labels after fine-tuning. [14], [5], [13], [25] implemented deep convolutional neural networks to rate the photo aesthetics, and achieved satisfying results. Our method is related to [14] and [5], but there are several differences with them. Unlike these methods, we train our CNN on the base of AlexNet model and VGGNet model. In other words, we do not train from scratch or simply impose deep learning as a feature extraction layer.

On a related topic, [9] described an approach to predicting style of images. Their experimental results demonstrated that mid-level features derived from object datasets are generic for style recognition.

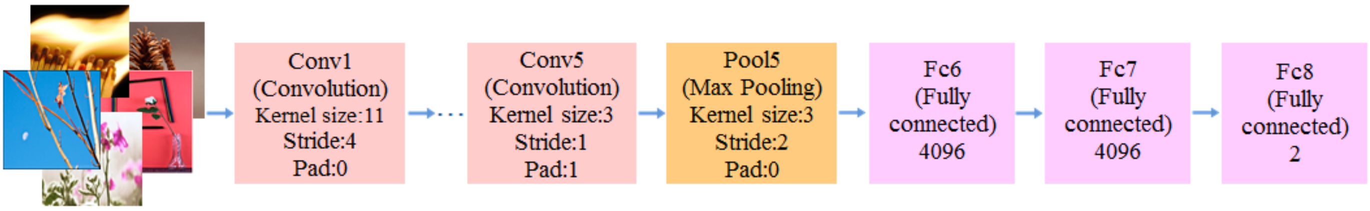


Figure 2. Finetuning a Convolutional Neural Network for visual aesthetics classification. In this example, we construct the model based on AlexNet. A binary label is assigned for each image. All the input images are normalized to the size of 256 × 256.

III. Visual aesthetics classification using CNNs

Our goal is to predict aesthetics by finetuning deep convolutional neural networks. We treat image aesthetic assessment as a binary classification task, and construct our single column CNN network on base of the AlexNet and VGGNet models.

A. Label generation

Image aesthetic datasets usually have scores from multiple human subjects. For example, the labels can range from 1-10, with 10 being excellent quality. In visual aesthetic classification, the scores are usually averaged, then the images are divided into two groups: low aesthetic quality (bad) and high aesthetic quality (good) images according to their average aesthetic scores. As a convention, images with average scores less than 5 are refers to as bad images, those with mean scores more than or equal to 5 are good ones. This makes the visual aesthetics classification a binary classification problem.

There are other criteria to pose the problem. In [14], the authors tested the case where the medium quality images were excluded. That is, they used the images with score larger than $5+\delta$ as positive, and score smaller $5-\delta$ as negative. In [5], the authors selected the top and bottom 10% images as positive and negative samples, respectively. We compared our methods to all these partitions in this paper.

B. Finetuning CNNs

Training CNNs on large scale image datasets gives us generic features for image classification, which can be a very good starting point towards many other computer vision tasks. Inspired by the recent work on finetuning CNNs, we attempt to learning features for aesthetic classification from pretrained models.

For the purpose of a comprehensive study, we chose two different CNNs, namely, AlexNet and VGGNet-16, respectively. Both networks achieved state of the art performances at different times, but their difference is also obvious. For example, VGGNet is deeper than the AlexNet, which means it may take more time to (re-)train the network.

1) AlexNet and VGGNet Models: AlexNet [12] contains five convolutional layers and three fully-connected layers. Output of every convolutional and fully-connected layers is modeled by Rectified Liner Unit (Relu). Response-normalization layers follow the first and second convolutional layers. Max-pooling layers follow both normalization layers and the fifth convolutional layer. The output of the last layer is followed by 1000-dimension classifier which classify 1000 different objects.

VGGNet is a very deep convolutional network which secured the first and the second places in the localisation and classification tasks respectively of ILSVRC-2014 [20]. It increased the depth to 16-19 weight layers. To reduce the number of parameters in such a very deep networks, very small filters (3 × 3) are used in all convolutional layers.

2) Finetuning pretrained networks: We use AlexNet as an example in this section to illustrate our finetuning steps.

We construct our network using AlexNet. As indicated in Figure 2, we first replace the last layer by a 2-class softmax classifier, because our goal is to classify images into bad or good aesthetics. Then, we set a larger learning rate for the last layer and smaller for the previous layers, and retrain the network.

Specifically, we did not use validation set in the training. Instead, we report the number at $10,000$ iteration. The learning rate was set to 0.001 for the last layer, which was initialized using the Xavier method. The learning rate was set to 0.0001 for other layers, which was initialized from pretrained models. Computational time is approx 30 minutes for AlexNet and 1 hour for VGGNet in our hardware platform.

C. Training Procedure

For each label generation setting (see Sec. III-A), we further generated global view and local view of each image respectively, and compared the performances when they were used to train our convolutional network models.

The global view refers to using the whole image as input. To get the global view images, we simply resize the images to 256 × 256.

There are different ways to obtain the local views. Randomly selecting a patch or using the center piece are two popular approaches. We selected the central patches to get the 256 × 256 local view images, because we think the central area includes most aesthetic features of photos.

In each setting, the images were separated to training and test sets, and fed to the network randomly. Then the convolutional networks were finetuned to learn the aesthetic features automatically. We set the batch size to 50 for finetuning AlexNet, and 20 for VGGNet. We stop training when the number of iterations is not more than 10,000 since the performance of the network becomes stable, which is shown in Sec. IV.

IV. Experiments

We use deep learning framework Caffe [7] to train our deep convolutional neural network, and to classify high and low aesthetic quality photographs. We tested our method on two datasets: CUHKPQ [24] and AVA analysis (AVA) [17].

The AlexNet and VGGNet were finetuned on both the local view and the global view. In total, four variations of the finetuning were studied.

A. Datasets

1) CUHKPQ: CUHKPQ consists of 17960 photos. It was released in [24] and can be downloaded from mmlab.ie.cuhk.edu.hk/CUHKPQ/Dataset.htm. The photos are divided into seven categories according to their themes. The quality of each photo is judged by ten independent viewers and labeled as high quality or low quality only if eight out of ten viewers have the same opinions.

In our experiments, we followed [5] and created the CUHKPQ (small) dataset by evenly splitting the high and low quality photos of each categories, which resulted in 8,845 images for training and 8,845 images for testing. We further assembled the CUHKPQ (Large) dataset by incorporating additional image files available in the download pages (which was not used in the [5]). This gives us 14,845 images for training and 14,845 images for testing.

2) AVA: AVA is a collection of images and meta-data derived from www.dpchallenge.com. It contains over 250,000 images along with a large number of aesthetic scores for each image. As reported in [17], Such scores have a high intrinsic value because there is an average of 210 amateur and professionals vote for each image.

We followed [5] and created the AVA (small) dataset by selecting the top and bottom 10%, which resulted in 19,308 images for training and 19,308 images for testing. We also used [14] and assembled the AVA (Large) dataset, which gives us 230,000 images for training and 20,000 images for testing.

B. Results of CUHKPQ dataset

We show the results on the CUHKPQ dataset in Table I and II.

Table I
Accuracy of CUHKPQ dataset (Large)

Local view, AlexNet	83.24%
Local view, VGGNet	87.90%
Global view, AlexNet	86.72%
Global view, VGGNet	91.43%

Table II
Accuracy of CUHKPQ dataset (Small)

Local view, AlexNet	88.56%
Local view, VGGNet	91.47%
Global view, AlexNet	91.20%
Global view, VGGNet	93.52%
method in [5]	91.93%

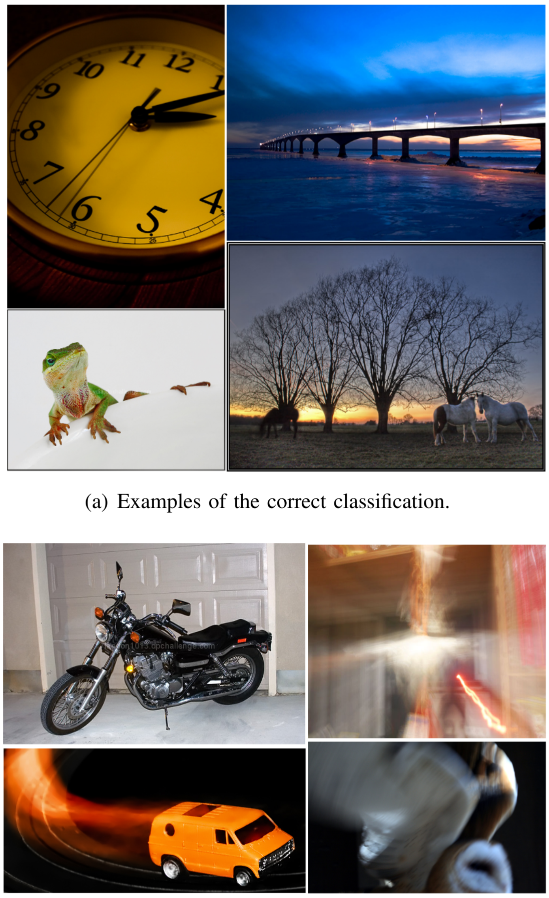
As shown in table I, our accuracy of aesthetic categorization is 91.43% on the large CUHKPQ dataset and 93.52% on the small set. Compared to [5], which tested their methods on the small set, we achieved approximately 1.6% better than theirs (91.93%).

The comparison between two pretrained networks is also meaningful. In all four experiments, VGGNet consistently outperformed AlexNet by a reasonable margin. However, as we show in Sec. III-B, AlexNet is much faster to reach satisfying results.

The difference between local view and global view suggests that we can simply use the global view, which contains the context information, to achieve better results.

C. Results of AVA dataset

1) Examples: We first present some examples in the AVA dataset. To simplify the visualization, we only show the results when the quality prediction is "good".



(a) Examples of the correct classification.

(b) Examples of the incorrect classification.

Figure 3. Examples of the classification.

Figure 3(a) shows the results when the input were labeled as "good" and correctly labeled. Figure 3(b) shows the failure examples, when the input were labeled as "bad" but were misclassified as "good" quality.

2) Statistics: As shown in Table III and Table IV, our accuracy of aesthetic categorization is 79.00% on the large AVA dataset and 85.41% on the small set, when global view is used. However, it is interesting to see that fine tuning VGGNet using local view has the highest accuracy (82%). This could be due to the characteristics of the large dataset.

Compared to [14], which tested their methods on the large set, we achieved approximately 8% better than theirs (VGGNet on Global View). [14] suggested the local view may be a better input than the global view. In our study, we suggested global view is a better source. The inconsistency between our result and theirs lies in the fact that we finetune our network from a pretrained model, which was trained from a more diversified data source. As a result, we can achieve better result by exploiting rich features in other datasets.

The comparison between two pretrained networks is consistent with our experiments on the CUHKPQ dataset.

Table III
ACCURACY OF AVA DATASET (LARGE)

Local view, AlexNet	74.99%
Local view, VGGNet	82.00%
Local view, method in [14]	71.20%
Global view, AlexNet	74.36%
Global view, VGGNet	79.00%
Global view, method in [14]	67.79%

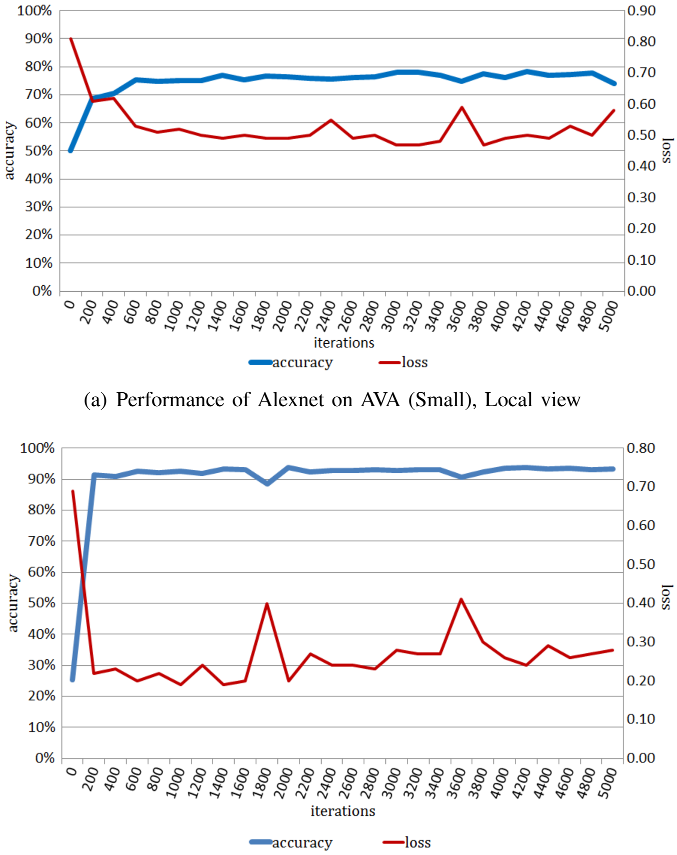
Table IV
ACCURACY OF AVA DATASET (SMALL)

Local view, AlexNet	78.72%
Local view, VGGNet	81.13%
Global view, AlexNet	83.24%
Global view, VGGNet	85.41%
method in [5]	83.52%

D. *Finetuning over time*

Compared to training from scratch, another significant advantage is that finetuning can achieve good results in a short time. In this section we show the performance of finetuning when the number of iteration increases.

Figure 4 shows one instance of the accuracy and the loss changes at different iterations for AlexNet and VGGNet, respectively. One can see that in each figure the loss decreases in just a few hundred iterations, and the accuracy reaches to a satisfying level at the same time. This makes it practical to retrain a CNN for a specific task like Visual Aesthetic Classification, and the overhead is minimum. We also observe that the networks may overfit if we continue the retraining. Therefore, it is sufficient to stop the finetuning in a small number of iteration. In our experiments, we found that 5,000 is sufficient to reach a satisfactory result, and 10,000 provides more stable models.



(a) Performance of Alexnet on AVA (Small), Local view

(b) Performance of VGGNet on CUHKPG (small), Global View

Figure 4. Preformance of Alexnet and VGGNet at different iterations.

V. CONCLUSION

We proposed to finetune CNNs for visual aesthetics classification, a challenging task in computer vision. We chose pretrained AlexNet model [12] and the 16-layer VGGNet model [20] to exploit the information from much larger scale and more diversified image datasets for improving our task. In our experiments, we tested our methods on AVA and CUHKPQ, using different configurations of training-testing partitions. We also compared the performance using both the local information and context information. Experiments suggest that our method outperforms the state of the art methods.

ACKNOWLEDGMENTS

NICTA is funded by the Australian Government as represented by the Department of Communications and the Australian Research Council.

This work was supported by the Australian Research Council (ARC) under Grant DP150104645. This work was also supported by “Jiangsu Overseas Research & Training Program for University Prominent Young & Middle-aged Teachers and Presidents” of Jiangsu Provincial Department of Education, China.

REFERENCES

[1] Östen Axelsson. Towards a psychology of photography: Dimensions underlying aesthetic appeal of photographs 1, 2, 3. *Perceptual and Motor Skills*, 105(2):411–434, 2007.

[2] Tunc Ozan Aydin, Aljoscha Smolic, and Markus Gross. Automated aesthetic analysis of photographic images. *Visualization and Computer Graphics, IEEE Transactions on*, 21(1):31–42, 2015.

[3] Christel Chamaret and Fabrice Urban. No-reference harmony-guided quality assessment. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 961–967. IEEE, 2013.

[4] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *Computer Vision–ECCV 2006*, pages 288–301. Springer, 2006.

[5] Zhe Dong, Xu Shen, Houqiang Li, and Xinmei Tian. Photo quality assessment with dcnn that understands image well. In *MultiMedia Modeling*, pages 524–535. Springer, 2015.

[6] Yee Whye Teh Geoffrey E Hinton, Simon Osindero. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[7] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[8] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 28(5):94–115, 2011.

[9] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *Proceedings of the British Machine Vision Conference. BMVA Press*, 2014.

[10] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 419–426. IEEE, 2006.

[11] Shehroz S Khan and Daniel Vogel. Evaluating visual aesthetics in photographic portraiture. In *Proceedings of the Eighth Annual Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging*, pages 55–62. Eurographics Association, 2012.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[13] Xin Lu, Zhe Lin, Hailin Jin, and Jianchao Yang. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 17(11):2021–2034, 2015.

[14] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia*, pages 457–466. ACM, 2014.

[15] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1784–1791. IEEE, 2011.

[16] E. Mavridaki and V. Mezaris. A comprehensive aesthetic quality assessment method for natural images using basic rules of photography. In *IEEE International Conference on Image Processing*, 2015.

[17] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2408–2415. IEEE, 2012.

[18] Masashi Nishiyama, Takahiro Okabe, Imari Sato, and Yoichi Sato. Aesthetic quality classification of photographs based on color harmony. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 33–40. IEEE, 2011.

[19] Gabriele Peters. Aesthetic primitives of images for visualization. In *Information Visualization, 2007. IV'07. 11th International Conference*, pages 316–325. IEEE, 2007.

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[21] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *null*, page 1470, 2003.

[22] Hsiao-Hang Su, Tse-Wei Chen, Chieh-Chi Kao, Winston H Hsu, and Shao-Yi Chien. Scenic photo quality assessment with bag of aesthetics-preserving features. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1213–1216. ACM, 2011.

[23] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[24] Xiaoou Tang, Wei Luo, and Xiaogang Wang. Content-based photo quality assessment. In *Computer Vision, IEEE International Conference on*, pages 2206–2213, 2011.

[25] Xinmei Tian, Zhe Dong, Kuiyuan Yang, and Tao Mei. Query-dependent aesthetic model with deep learning for photo quality assessment. *IEEE Transactions on Multimedia*, 17(11):2035–2048, 2015.

[26] Geoffrey Hinton Yann LeCun, Yoshua Bengio. Deep learning. *Nature*, 521:436–444, 2015.