Deep Image Aesthetics Classification using Inception Modules and Fine-tuning Connected Layer

Xin Jin[1,*], Jingying Chi[2], Siwei Peng[2], Yulu Tian[1], Chaochen Ye[1] and Xiaodong Li[1,*]
[1]Beijing Electronic Science and Technology Institute, Beijing 100070, China
[2]Beijing University of Chemical and Technology, Beijing 100029, China
Corresponding Authors: {jinxin,lxd}@besti.edu.cn

***Abstract*—In this paper we investigate the image aesthetics classification problem, aka, automatically classifying an image into low or high aesthetic quality, which is quite a challenging problem beyond image recognition. Deep convolutional neural network (DCNN) methods have recently shown promising results for image aesthetics assessment. Currently, a powerful inception module is proposed which shows very high performance in object classification. However, the inception module has not been taken into consideration for the image aesthetics assessment problem. In this paper, we propose a novel DCNN structure codenamed ILGNet for image aesthetics classification, which introduces the Inception module and connects intermediate Local layers to the Global layer for the output. Besides, we use a pre-trained image classification CNN called GoogLeNet on the ImageNet dataset and fine tune our connected local and global layer on the large scale aesthetics assessment AVA dataset [1]. The experimental results show that the proposed ILGNet outperforms the state of the art results in image aesthetics assessment in the AVA benchmark.**

I. Introduction

In practice, mastering the technical aspects of shooting good photos is an acquired skill that takes years of vigilant observation to learn. However, people often can easily distinguish whether an image is beautiful or not. As shown in 1, most people will prefer the left images as they are more beautiful than those in the right.

Nowadays, facilitate mobile devices, social networks, and cloud storages make the fast increasing of the amount of images of home users or in the Internet. Thus the ability of automatically classify an image to low or high aesthetic quality can be used in various scenarios, such as follows.

- To return Internet image search results with high aesthetic quality;
- Today, people often make crazy shooting in daily life using their mobile phones. After that, they often struggle to select good photos from thousands of photos for sharing in their social network. Thus, the image aesthetics classification algorithm can help them to automatically select most beautiful images for sharing;
- Image aesthetics classification also helps to develop new image beautification tools to make images look better [2];
- The vast amount of work from graphic, architecture, industry, and fashion design can be automatically classified to low or high quality.

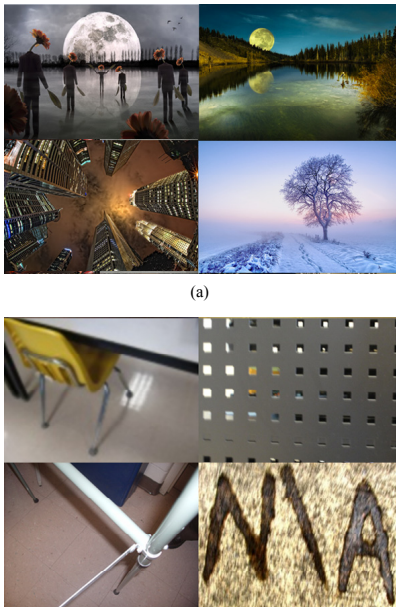

Fig. 1. For most people, they may consider that the left images in (a) are more attractive than those in (b). Images are from the AVA dataset. [1]

Subjective Image Aesthetic Quality Assessment (IAQA) is still challenging [2], which aim to automatically classify a image into low or high aesthetic quality or giving a numerical assessment of the aesthetic quality. The challenges mainly come from the followings.

- the large intra class difference of high or low aesthetics;
- plenty of low level features and high level aesthetics rules;
- the subjective evaluation of human rating.

Thus, this problem has becoming a hot topic in the communities of Computer Cision (CV), Computational Aesthetics (CA) and Computational Photography (CP). In early work, various hand-crafted aesthetic features (i.e. aesthetic rule based features) are designed and connected with a machine classification or regression [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. Another line is to use generic image description features [21], [22], [23], [24]. After that, deep learning methods, which have shown

great success in various computer vision tasks, have recently been used to extract effective aesthetics features [25], [26], [27], [28], [29], [30], [2], [31], [32], [33] .

Recently, an efficient deep neural network architecture for computer vision, codenamed Inception is proposed by Google [34]. The inception module derives its name from the Network in network paper by Lin et al [35] in conjunction with the famous *we need to go deeper* internet meme [36]. In general, one can view the inception module as a logical culmination of [34] while taking inspiration and guidance from the theoretical work by Arora et al [37]. The benefits of the architecture are experimentally verified on the ILSVRC 2014 classification and detection challenges, where it significantly outperforms the state of the art before the year 2015. However, to the best of our knowledge, little attention has been paid to use inception for image aesthetic quality assessment in current literatures.

In this paper, we introduce the inception module into image aesthetics classification. We build a novel Deep convolutional neural network, codenamed ILGNet (I: Inception, L: Local, G: Global) using multiple inception modules. Recent work [38] [34] shows value in directly connecting intermediate layers to the output. Thus, we connect the layers of local features to the layer of global features. The network is 13 layers deep when counting only layers with parameters (or 17 layers if we also count pooling). Firstly, we train our ILGNet on the ImageNet [39], which is the largest available image dataset for 1000 categories object classification. Then we approximately fixed the inception layers and fine tune the connected layer contains global and local features on the largest available image aesthetics dataset, the AVA dataset [1]. The experimental results on the AVA dataset [1] outperform the state of the art in image aesthetics classification.

II. Related Work

In this section we briefly investigate the related work of our image aesthetics classification: The objective image quality assessment, the image aesthetic quality assessment using hand-crafted features, the deep image aesthetic quality assessment.

A. Objective Image Quality Assessment

Objective image quality assessment aim to evaluate image quality distorted by imaging, transmission, and compression. They detect and measure various distortions including blocking,ringing, mosaic patterns, blur, noise, ghosting, jerkiness,smearing, etc [2]. These low-level distortion measurement-based metrics can not well model human perception of the image aesthetic quality.

B. Aesthetic Quality Assessment with Hand-crafted Features

Subjective image aesthetic quality assessment aim to automatically classify a image into low or high aesthetic quality or giving a numerical assessment of the aesthetic quality. In this area, researchers usually following three standard steps.

- They collect a dataset of images and manually separate them into two subjects: (1) the images with high aesthetic quality, labelled as *good* or *high*, (2) the ones with low aesthetic quality, labelled as *bad* or *low*. Some work pick up some of the images and make psychological experiments to obtain numerical assessment of the aesthetic quality of images.
- They design various aesthetics orientation features such as rule of third, visual balance, rule of simplicity [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. In another way, they use generic image features for object recognition, such as low level image features[21], Fisher Vector [22] and bag of visual words [23], [24] to predict image aesthetics.
- They use machine learning tools such as SVM, Adaboost, and Random Forest to train a classifier on the collected datasets to automatically predict the aesthetic label of image (high or low, good or bad). They regress the hand-crafted design features to the human evaluated scores to predict the numerical assessment results of the image aesthetic quality.

C. Deep Image Aesthetic Quality Assessment

Recently, deep learning methods have shown great success in various computer vision tasks, such as object recognition, object detection, and image classification [34], [40]. Deep learning methods, such as deep convolutional neural network and deep belief network, have also been applied to image aesthetics assessment and have significantly improve the prediction precision against non-deep methods [25], [26], [27], [28], [29], [30], [2], [31], [32], [33]. Most of the architectures follow the AlexNet [41], which is an 8 layers network with 5 convolutional layers and 3 full-connected layers. Although good performance they obtain, inspired by the recent achievement by Google in the ILSVRC challenge, we should go deeper with multiple inception modules. Besides, recent work [38] [34] shows value in directly connecting intermediate layers to the output. Thus we change our network by connecting the intermediate local feature layers to the global feature layer.

III. Image Aesthetics Classification via ILGNet

In this section we will describe the details of our proposed ILGNet. As shown in 2, our network is 13 layers deep when counting only layers with parameters or 17 layers if we also count pooling. Three inception layers and one pre-treatment layer are involved. We connect the two intermediate layers of local features to the layer of global features to form a concat layer of 1024 dimension, following a full connected layer. The output layer is 1 dimension which directly give the classification result of low or high aesthetic quality.

A. The Inception Module

The main idea of the Inception architecture is to consider how an optimal local sparse structure of a convolutional vision network can be approximated and covered by readily available dense components. As shown in Fig. 3, in order to avoid patch-alignment issues, the incarnations of the Inception architecture are restricted to filter sizes $1 * 1$, $3 * 3$ and $5 * 5$. As these

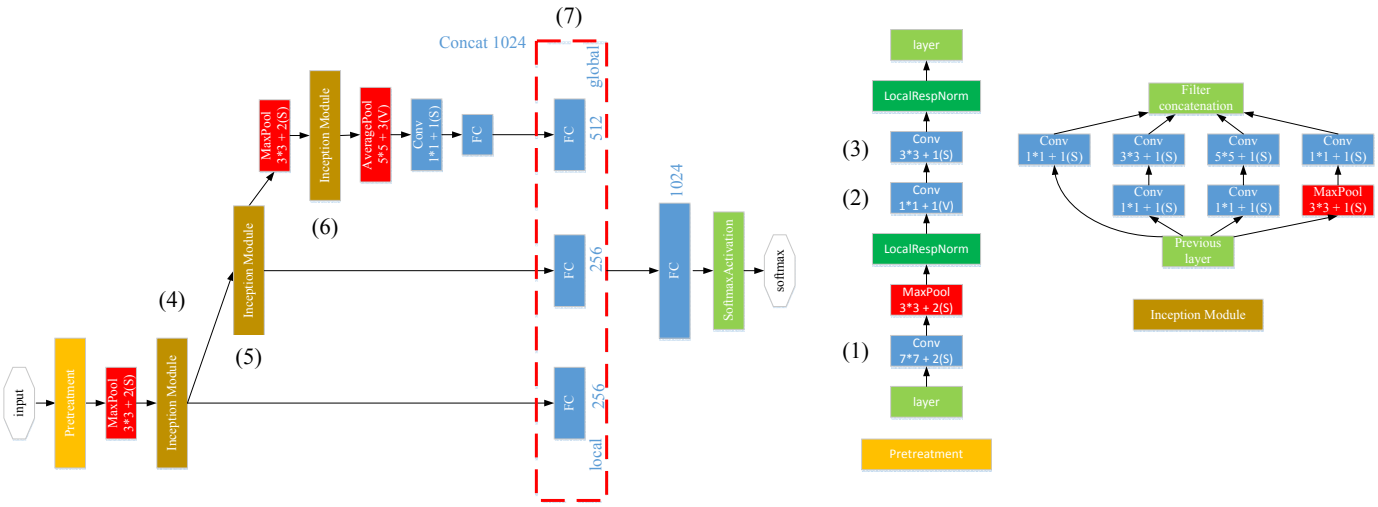

Fig. 2. The architecture of the proposed ILGNet: Inception with connected Local and Global layers. We use one pre-treatment layer and three inception layers. The first two inception layers extract local features and the last one extracts global features. Recent work [38] [34] shows value in directly connecting intermediate layers to the output. Thus, we connect the two layers of local features to the layer of global features to form a concat layer of 1024 dimension to a full connected layer. The output layer is 1 dimension which indicate low or high aesthetic quality. The network is 13 layers deep when counting only layers with parameters (or 17 layers if we also count pooling). The labels (1)-(7) are used for the visualization in section IV.

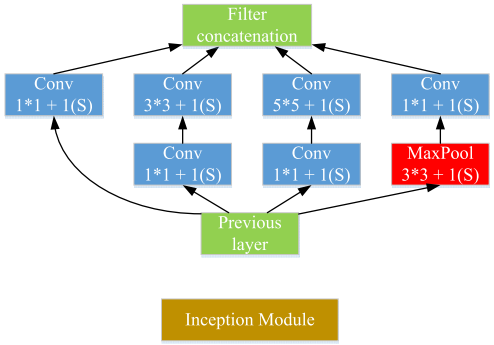

Fig. 3. The detail of the inception module [34].

Inception modules are stacked on top of each other, their output correlation statistics are bound to vary: as features of higher abstraction are captured by higher layers, their spatial concentration is expected to decrease. The ratio of $3*3$ and $5*5$ convolutions should increase as we move to higher layers. the stride is 1.

In general, an Inception network is a network consisting of modules of the above type stacked upon each other, with occasional max-pooling layers with stride 2 to halve the resolution of the grid. For technical reasons (memory efficiency during training), it seemed beneficial to start using Inception modules only at higher layers while keeping the lower layers in traditional convolutional fashion. This is not strictly necessary, simply reflecting some infrastructural inefficiencies in the implementation [34].

B. The ILGNet for Aesthetics Prediction

All the convolutions, including those inside the Inception modules, use rectified linear activation. The size of the receptive field in our network is $224 * 224$ in the RGB color space with zero mean. All these reduction/ projection layers use rectified linear activation as well [34].

The first and the second inception layers are considered to extract local image features. The last inception layer is considered to extract global image features after two max pooling and one average pooling. Then, we connect the output of the first two inception layers (256 dimension for each) and last inception layer (512 dimension) to form a 1024 dimension concat layer. This contact layer is followed by a full connected layer with the same dimension. the output of our ILGNet is bypass a softmax layer to a binary output, which indicate low or high aesthetic quality of an image.

Firstly, we train our ILGNet on the ImageNet [39], which is the largest available image dataset for 1000 categories object classification. Then we approximately fixed the inception layers and fine tune the connected layer contains global and local features on the largest available image aesthetics dataset, the AVA dataset [1].

IV. Experiments

In this section, we report the experimental results to verify the effectiveness of our proposed ILGNet when dealing with image aesthetics classification. It will be compared with several state-of-the-art methods. Most of them are based on deep neural networks. All the experiments are conducted on the large scale and reliable public datasets AVA, which is specifically designed for the research of photo quality assessment [1].

A. Dataset

1) The ImageNet Dataset: The ILSVRC 2014 classification challenge involves the task of classifying the image into one

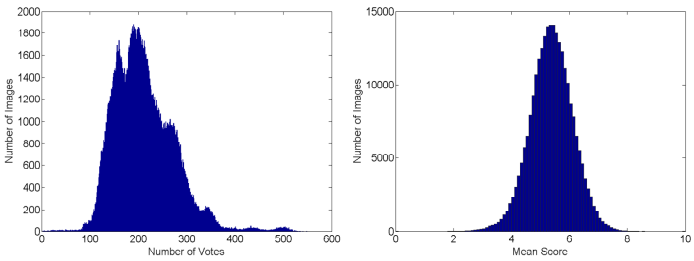

Fig. 4. The histogram/distribution of the mean scores and the number of votes per image in the AVA dataset.

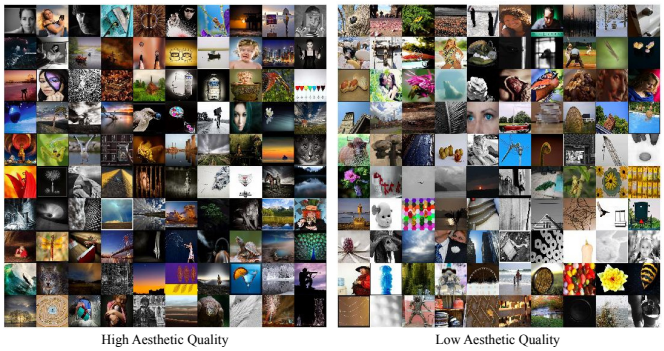

Fig. 5. An embedding of the AVA dataset. The left and right part are the high (mean score above 5) and low aesthetic quality (mean score below 5), respectively. Note that, it is not easy to make a statistical description of the difference between the two sub-datasets by human.

of 1000 leaf-node categories in the Imagenet hierarchy. There are about 1.2 million images for training, 50,000 for validation and 100,000 images for testing. Each image is associated with one ground truth category. Firstly, we train our ILGNet on the 1.2 million training images from ILSVRC 2014 classification challenge for 1000 categories.

2) The AVA Dataset: Aesthetic Visual Analysis (AVA) [1] is a large dataset formed by more than 250 thousands of images [25]. This database is specifically constructed for the purpose of learning more about image aesthetics. All those images are directly downloaded from the DPChallenge.com. For each image in AVA, there is an associated distribution of scores (0-10) voted by different viewers. As reported in [1], the number of votes that per image gets is ranged in 78-549, with an average of 210, as shown in Fig. 4. Fig. 5 shows an embedding of the AVA dataset.

B. Classification Results

For a fair comparison, we adopted same strategy to construct two sub dataset of AVA as the previous work.

- AVA1: We chose the score of 5 as the boundary to divide the dataset into high quality class and low quality class. In this way, there are 74,673 images in low quality and 180,856 images in high quality. the training and test sets contain 235,599 and 19,930 images respectively [1], [30], [33], [32], [28], [27], [2].

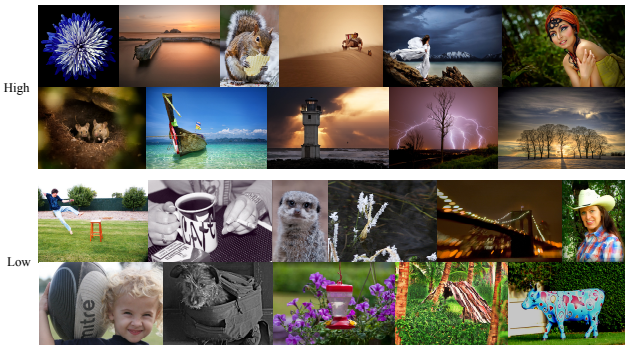

Fig. 6. The images classified by our ILGNet to high (top) or low (bottom) aesthetic quality.

TABLE I
THE CLASSIFICATION ACCURACY IN AVA1 DATASET.

Methods	Accuracy
MurrayCVPR2012 [1]	67.0%
WangSP2016 [30]	76.94%
WangCORR2016 [33]	76.8%
KongECCV2016 [32]	77.33%
LuTMM2015 [28]	74.46%
LuICCV2015 [27]	75.41%
MaiCVPR2016 [2]	77.1%
Our ILGNet	**79.25%**

TABLE II
THE CLASSIFICATION ACCURACY IN AVA2 DATASET.

Methods	Accuracy
LuoECCV2008 [5]	61.49%
LoICPR2012 [42]	68.13%
DattaECCV2006 [3]	68.67%
KeCVPR2006 [4]	71.06%
MarchesottiICCV2011 [22]	68.55%
DongNC2015 [29]	78.92%
DongMMM2015 [43]	83.52%
WangSP2016 [30]	84.88%
Our ILGNet	**85.62%**

- AVA2: to increase the gap between images with high aesthetic quality and images with low aesthetic quality, we firstly sort all images by their mean scores. Then we pick out the top 10% images as good and the bottom 10% images as bad. Thus, we select 51,106 images form the AVA dataset. And all images are evenly and randomly divided into training set and test set, which contains 25,553 images respectively [5], [42], [3], [4], [22], [29], [43], [30].

The sample classification results using our ILGNet is shown in Fig. 6. Differences between low-aesthetic images and high-aesthetic images heavily lie in the amount of textures and complexity of the entire image [28].

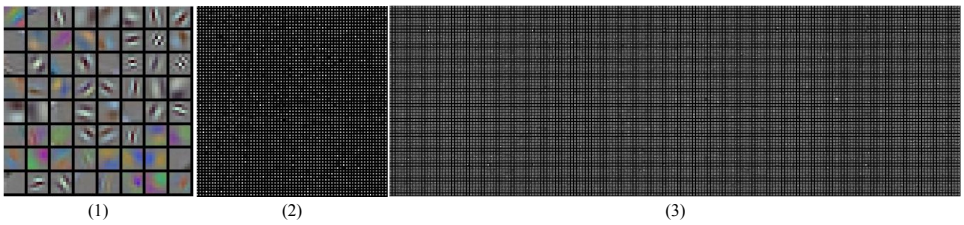

Fig. 7. The visualization results of the the weights of the first three convolutional layers. The labels of (1), (2), (3) correspond to the same labels in Fig. 2.

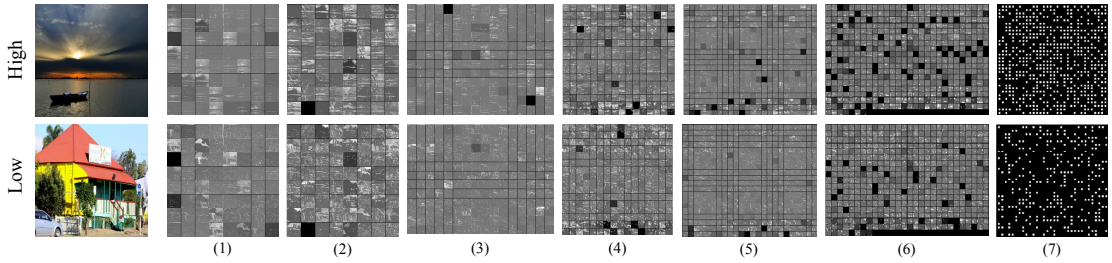

Fig. 8. The visualization results of the the weights of the features extracted by our ILGNet in important layers for images with high (top) and low (bottom) labels. The labels of (1)-(7) correspond to the same labels in Fig. 2.

We present the experimental results in the AVA1 dataset in Table I. It can be observed that our ILGNet outperforms the state of the art DCNN architectures with the accuracy 79.25%. The best performance obtained by the methods based on hand-crafted features is 67.0% [1], which is worse than the DCNN features.

The classification accuracy in the AVA2 dataset is shown in Table II. Our ILGNet outperforms the state of the art DCNN architectures with the accurcy 85.62%. The increasing of the gap between the high and low images significantly increases the classification accuracy. The best performance obtained by the methods based on hand-crafted features is 68.55% [22], which is still worse than DCNN architectures.

C. Visualization

We visualize our learned ILGNet in the aspects of the network weights and the features.

1) The Network Weights: The ILGNet is 13 layers deep when counting only layers with parameters (or 17 layers if we also count pooling). We visualize the weights of the first three convolutional layers, as shown in Fig. 7. We first train the network in the ImageNet dataset. Thus the learned weights of the first three convolutional layers can extract the generic low level image features.

2) The Features: We visualize the extracted features by our LGNet from images with high and low aesthetic quality. As shown in Fig. 8, our ILGNet can extract features from the low level to high level. The last feature maps shown the features extracted by the connected layer of local and global feature extractors, which are binary pattens.

V. Conclusion and Discussion

In this paper, we propose a novel DCNN to predict the aesthetic label of low or high for images, codenamed ILGNet, which introduces multiple power inception modules and a connected local and global layer. We first train our ILGNet on the ImageNet [39]. Then we approximately fixed the inception layers and fine tune the connected layer on the AVA dataset [1]. This architecture goes in deeper than current DCNN used for image aesthetic quality assessment and outperforms the state of the art in the largest aesthetic image dataset: the AVA dataset with both two strategies of the dataset partition. In the future work, we will introduce more domain knowledge in this field into the design of the DCNN for image aesthetic quality assessment and try to make the architecture itself *learnable* in the future.

VI. Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Grant NO.61402021, 61402023), the Science and Technology Project of the State Archives Administrator (Grant NO.2015-B-10), the open funding project of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (Grant NO. BUAA-VR-16KF-09), and the Fundamental Research Funds for the Central Universities (NO. 2014GCYY02, 2014GCYY04, 2016LG03, 2016LG04).

References

[1] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, 2012, pp. 2408–2415.

[2] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part III*, 2006, pp. 288–301.

[4] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, 2006, pp. 419–426.

[5] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III*, 2008, pp. 386–399.

[6] C. Li and T. Chen, "Aesthetic visual quality assessment of paintings," *J. Sel. Topics Signal Processing*, vol. 3, no. 2, pp. 236–252, 2009.

[7] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, 2010, pp. 271–280.

[8] W. Jiang, A. C. Loui, and C. D. Cerosaletti, "Automatic aesthetic value assessment in photographic images," in *Proceedings of the 2010 IEEE International Conference on Multimedia and Expo, ICME 2010, 19-23 July 2010, Singapore*, 2010, pp. 920–925.

[9] C. Li, A. C. Gallagher, A. C. Loui, and T. Chen, "Aesthetic quality assessment of consumer photos with faces," in *Proceedings of the International Conference on Image Processing, ICIP 2010, September 26-29, Hong Kong, China*, 2010, pp. 3221–3224.

[10] X. Jin, M. Zhao, X. Chen, Q. Zhao, and S. C. Zhu, "Learning artistic lighting template from portrait photographs," in *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*, 2010, pp. 101–114.

[11] D. Gray, K. Yu, W. Xu, and Y. Gong, "Predicting facial beauty without landmarks," in *Computer Vision - ECCV 2010 - 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI*, 2010, pp. 434–447.

[12] X. Chen, X. Jin, H. Wu, and Q. Zhao, "Learning templates for artistic portrait lighting analysis," *IEEE Trans. Image Processing*, vol. 24, no. 2, pp. 608–618, 2015.

[13] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, 2011, pp. 1657–1664.

[14] D. Joshi, R. Datta, E. A. Fedorovskaya, Q. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Process. Mag.*, vol. 28, no. 5, pp. 94–115, 2011.

[15] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, 2011, pp. 33–40.

[16] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, 2011, pp. 2206–2213.

[17] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[18] O. Wu, W. Hu, and J. Gao, "Learning to predict the perceived visual quality of photos," in *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, 2011, pp. 225–232.

[19] S. S. Khan and D. Vogel, "Evaluating visual aesthetics in photographic portraiture," in *Computational Aesthetics 2012: Eurographics Workshop on Computational Aesthetics, Annecy, France, 4-6 June 2012. Proceedings*, 2012, pp. 55–62.

[20] Y. Niu and F. Liu, "What makes a professional video? A computational aesthetics approach," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 22, no. 7, pp. 1037–1049, 2012.

[21] H. Tong, M. Li, H. Zhang, J. He, and C. Zhang, "Classification of digital photos taken by photographers or home users," in *Advances in Multimedia Information Processing - PCM 2004, 5th Pacific Rim Conference on Multimedia, Tokyo, Japan, November 30 - December 3, 2004, Proceedings, Part I*, 2004, pp. 198–205.

[22] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, 2011, pp. 1784–1791.

[23] H. Su, T. Chen, C. Kao, W. H. Hsu, and S. Chien, "Scenic photo quality assessment with bag of aesthetics-preserving features," in *Proceedings of the 19th International Conference on Multimedia 2011, Scottsdale, AZ, USA, November 28 - December 1, 2011*, 2011, pp. 1213–1216.

[24] ——, "Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features," *IEEE Trans. Multimedia*, vol. 14, no. 3-2, pp. 833–843, 2012.

[25] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, "Recognizing image style," in *British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5, 2014*, 2014.

[26] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "RAPID: rating pictorial aesthetics using deep learning," in *Proceedings of the ACM International Conference on Multimedia, MM'14, Orlando, FL, USA, November 03 - 07, 2014*, 2014, pp. 457–466.

[27] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 990–998.

[28] X. Lu, Z. L. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.

[29] Z. Dong and X. Tian, "Multi-level photo quality assessment with multi-view features," *Neurocomputing*, vol. 168, pp. 308–319, 2015.

[30] W. Wang, M. Zhao, L. Wang, J. Huang, C. Cai, and X. Xu, "A multi-scene deep learning model for image aesthetic evaluation," *Signal Processing: Image Communication*, pp. –, 2016.

[31] Y. Kao, R. He, and K. Huang, "Visual aesthetic quality assessment with multi-task deep learning," *CoRR*, vol. abs/1604.04970, 2016.

[32] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *European Conference on Computer Vision (ECCV)*, 2016.

[33] Z. Wang, F. Dolcos, D. Beck, S. Chang, and T. S. Huang, "Brain-inspired deep networks for image aesthetics assessment," *CoRR*, vol. abs/1601.04155, 2016.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 1–9.

[35] M. Lin, Q. Chen, and S. Yan, "Network in network," *CoRR*, vol. abs/1312.4400, 2013.

[36] "Know your meme: We need to go deeper," http://knowyourmeme.com/memes/we-need-to-go-deeper, 2013.

[37] S. Arora, A. Bhaskara, R. Ge, and T. Ma, "Provable bounds for learning some deep representations," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, 2014, pp. 584–592.

[38] M. Maire, S. X. Yu, and P. Perona, "Reconstructive sparse code transfer for contour detection and semantic labeling," in *Computer Vision - ACCV 2014 - 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part IV*, 2014, pp. 273–287.

[39] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 2009, pp. 248–255.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, 2012, pp. 1106–1114.

[42] K. Lo, K. Liu, and C. Chen, "Assessment of photo aesthetics with efficiency," in *Proceedings of the 21st International Conference on Pattern Recognition, ICPR 2012, Tsukuba, Japan, November 11-15, 2012*, 2012, pp. 2186–2189.

[43] Z. Dong, X. Shen, H. Li, and X. Tian, "Photo quality assessment with DCNN that understands image well," in *MultiMedia Modeling - 21st International Conference, MMM 2015, Sydney, NSW, Australia, January 5-7, 2015, Proceedings, Part II*, 2015, pp. 524–535.