IET Computer Vision

Special Issue: Visual Domain Adaptation and Generalisation

ILGNet: inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation

ISSN 1751-9632
Received on 22nd December 2017
Revised 26th April 2018
Accepted on 22nd May 2018
E-First on 19th July 2018
doi: 10.1049/iet-cvi.2018.5249
www.ietdl.org

Xin Jin[1,2], Le Wu[1], Xiaodong Li[1], Xiaokun Zhang[1], Jingying Chi[3], Siwei Peng[3], Shiming Ge[4] ✉, Geng Zhao[1], Shuying Li[5]

[1]*Department of Computer Science and Technology, Beijing Electronic Science and Technology Institute, Beijing 100070, People's Republic of China*
[2]*CETC Big Data Research Institute Co., Ltd., Guiyang 550018, People's Republic of China*
[3]*College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, People's Republic of China*
[4]*Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, People's Republic of China*
[5]*The 16th Institute, China Aerospace Science and Technology Corporation, Xi'an 710100, People's Republic of China*
✉ *E-mail: geshiming@iie.ac.cn*

Abstract: In this study, the authors address a challenging problem of aesthetic image classification, which is to label an input image as high- or low-aesthetic quality. We take both the local and global features of images into consideration. A novel deep convolutional neural network named ILGNet is proposed, which combines both the inception modules and a connected layer of both local and global features. The ILGnet is based on GoogLeNet. Thus, it is easy to use a pre-trained GoogLeNet for large-scale image classification problem and fine tune their connected layers on a large-scale database of aesthetic-related images: AVA, i.e. *domain adaptation*. The experiments reveal that their model achieves the state of the arts in AVA database. Both the training and testing speeds of their model are higher than those of the original GoogLeNet.

1 Introduction

Shooting good photographs needs years of practise for photographers. However, it is often easy for people to classify an image into high or low-aesthetic quality. As shown in Fig. 1, the left image is often considered as with higher-aesthetic quality than the right one.

Recently, smartphones, social networks and cloud computing boost the number of images in the public or private cloud. People need a better way to manage their photographs than ever before. An important ability of today's photograph management software is to automatically recommend good photographs from a large number of daily photographs. Besides, aesthetic quality assessment can be used in the following scenarios:

i. When you search images on the Internet, the aesthetic

Today, image aesthetic quality classification is still a challenging problem. Typically, the following reasons make it challenging:

- Two classes of high- and low-aesthetic qualities contain large intra-class differences.
- Many high-level aesthetic rules versus low-level image features.
- The subjective nature of human rating on aesthetic qualities of images.

Thus, people from computer vision, computational photography and computational aesthetics make this topic hot. In their early work, they design hand-crafted aesthetic image features, which are fed into a classification model or a regression model. Generic image features are also used in aesthetic quality classification. Today, deep convolutional neural networks are designed especially architecture with inception module shows its benefits. The performance was significantly raised in the classification and detection challenges. However, in current literature, inception modules have not been used in the aesthetic quality assessment to the best of our knowledge.

We propose to use inception modules for image aesthetics classification in this paper. A new deep convolutional neural network using inception modules with connected low and global features is proposed, which is called ILGNet. Connecting intermediate layers directly to the output layers has shown its value in recent work [5, 8]. In our ILGNet, the local feature layers are connected to the global feature layers. The ILGNet contains 13 layers with parameters and without counting pooling layers (4 layers). We use a pre-trained model on the ImageNet [9] as our initial model, which is trained for object classification of 1000 categories. Then, the inception modules are fixed and the connected local and global features layers are finely tuned on the aesthetic visual analysis (AVA) database, which is currently largest image aesthetics database [10]. We achieve the state of the art in the experiments on the AVA database [10]. Besides, the trained models and codes are available at github: https://github.com/BestiVictory/ILGnet.

The rest of this paper is organised as follows. In Section 2, we review the related work. In Section 3, we describe our proposed ILGNet in details. Then, the experimental settings, results and comparisons with state-of-the-art methods are presented in Section 4. Finally, we give a conclusion in Section 5.

2 Previous work

The related work of our task can be categorised into the traditional image quality assessment, the subjective image aesthetic quality assessment using hand-crafted features and deep learning.

2.1 Traditional image quality assessment

Traditional image quality assessment is to assess the objective image quality, which may be distorted or influenced during the imaging, compression and transmission. Distortions such as ringing, blur, ghosting, smearing, blocking, mosaic and jerkiness are measured [11]. The human perception of aesthetics cannot be well modelled by these low-level features and metrics.

2.2 Hand-crafted features for subjective image aesthetic quality assessment

Subjective image aesthetic quality assessment is to automatically distinguish an image to low- or high-aesthetic quality. Some of them can give a numerical assessment. They often contain the three steps as follows:

- A database of images is collected. Then they often manually label each with two labels: *good* for images with high-aesthetic

deep convolutional neural network have been used for image aesthetics assessment. The performance has been significantly improved compared with traditional methods [11, 35–44].

Most of the above work has the AlexNet architecture [45], which contains eight layers with five convolutional layers and three full-connected layers or visual geometry group (VGG) [46]. Inspired by the good performance of GoogLeNet in the ImageNet, which argues that deeper architectures enable to capture large receptive field. We can extract local image features and the global features of the image layout. Connecting intermediate layers directly to the output layers has shown its value in recent work [5, 8]. Both the local features and the global features can be extracted by inception modules. Thus, we change the GoogLeNet by connecting the intermediate local feature layers to the global feature layer.

3 ILGNet for image aesthetic quality classification

The details of the proposed ILGNet are described in this section. The ILGNet contains 13 layers with parameters and without counting pooling layers (4 layers). The network contains one pre-treatment layer and three inception modules. Two intermediate layers of local features are connected to a layer of global features, which makes a 1024 dimension concat layer. The output layer indicates the probability of low- or high-aesthetic quality. The basic ILGNet is built on the first 1/3 part of GoogLeNetV1 [5] and batch normalisation, which is an important feature of GoogLeNetV2 [47].

3.1 Inception module

The InceptionV1 module is proposed by GoogLeNetV1 [5]. The main ideas of the inception module are:

i. Convolution kernels with different sizes represent receptive fields with difference sizes. This design means fusing features of different scales.
ii. The kernel sizes are set to 1×1, 3×3 and 5×5 so as to align the features conveniently. The stride is 1. The pad is set to 0, 1, 2.
iii. The features extracted by the higher layer are increasingly abstract. The receptive field involved by each feature is larger. Thus, the ratio of 3×3 and 5×5 kernels should be increased.

After InceptionV1, Google proposed InceptionV2 and InceptionV3, which adopt factorisation of convolutions and improved normalisation. Then, InceptionV4 considered the residue network, which surpassed its ancestor GoogLeNet on the ImageNet benchmark.

3.2 Image aesthetic quality classification

The convolution layers inside ILGNet has rectified linear activation. The size of the input receptive field of ILGNet is

Table 1 Main training parameters of the Caffe package

Parameters	AVA1 ($\delta = 0$)	AVA1 ($\delta = 1$)	AVA2
base_ lr	0.0001	0.00001	0.00001
lr_ policy	'step'	'step'	'step'
stepsize	100,000	19,000	13,325
gamma	0.96	0.96	0.96
max_iter	475,000	760,000	533,000
momentum	0.9	0.9	0.9
weight_decay	0.0002	0.0002	0.0002

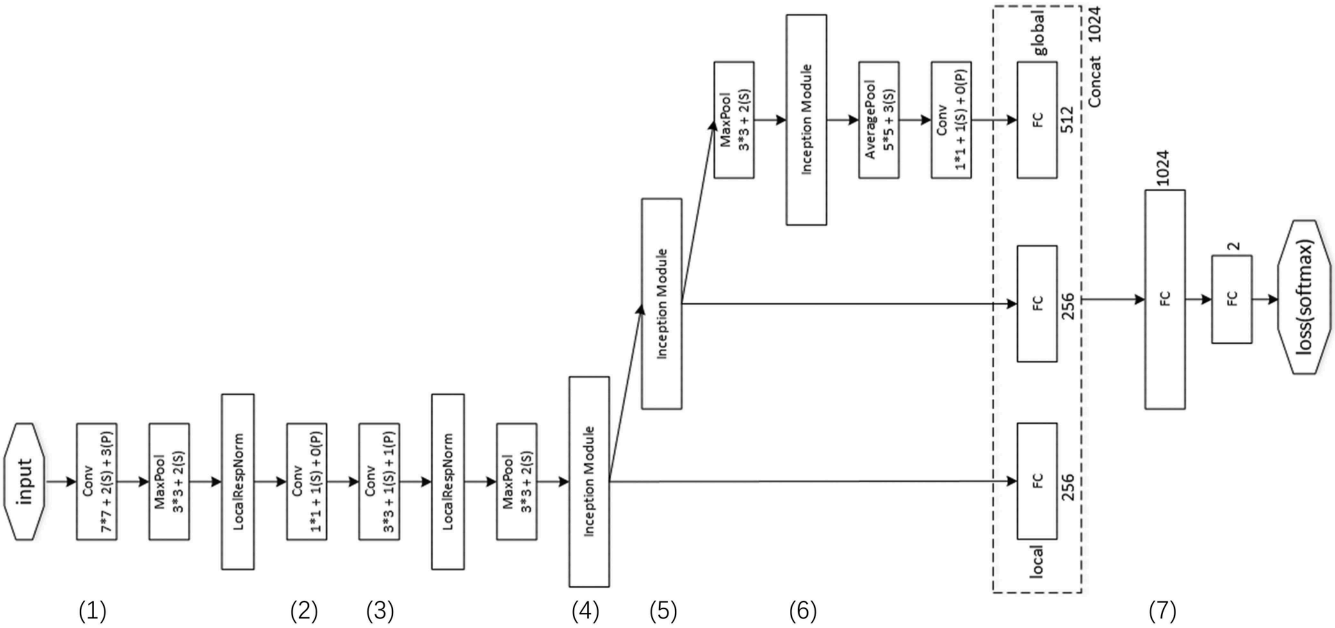


Fig. 2 *We use the trained ILGNet to label images with good or bad, which indicates high- or low-aesthetic quality, respectively*

them use deep convolutional neural networks. The main training parameters of the Caffe package [48] are listed in Table 1.

4.1 Database and comparison protocols

The aesthetic visual analysis database [10] is a list of image ids from DPChallenge.com, which is an online photography social network. There are total 255,529 photographs, each of which is rated by 78–549 persons, with an average of 210. The range of the scores rated by a human is 1–10. We use the same protocols to those of the previous work. They often use two sub-database of AVA:

- *AVA1*: The score of five are chosen as the threshold to

The original ILGNet is built on the first 1/3 of GoogLeNet V1, as shown in Fig. 3. We add batch normalisation (GoogLeNet V2 [5] features), which form our ILGNet-Inc.V1-BN. After that, we further built our ILGNet on the first 1/3 of recent GoogLeNet V3 [51] and V4 [52], which form our ILGNet-Inc.V3 and ILGNet-Inc.V4. The test results in the AVA1 database are shown in Table 2. Our ILGNet-Inc.V4 outperforms the other deep convolutional neural networks (DCNN)-based methods and achieve the state-of-the-art accuracy: 82.66%.

The above is the case of $\delta = 1$. Similar results are shown when $\delta = 1$. In the original test protocol [10], they set $\delta = 1$ in the training set, there are 7500 low-quality images and 45,000 high-quality images. For the testing images, they fix δ to 0, regardless of what δ is used for training. We have tested five network

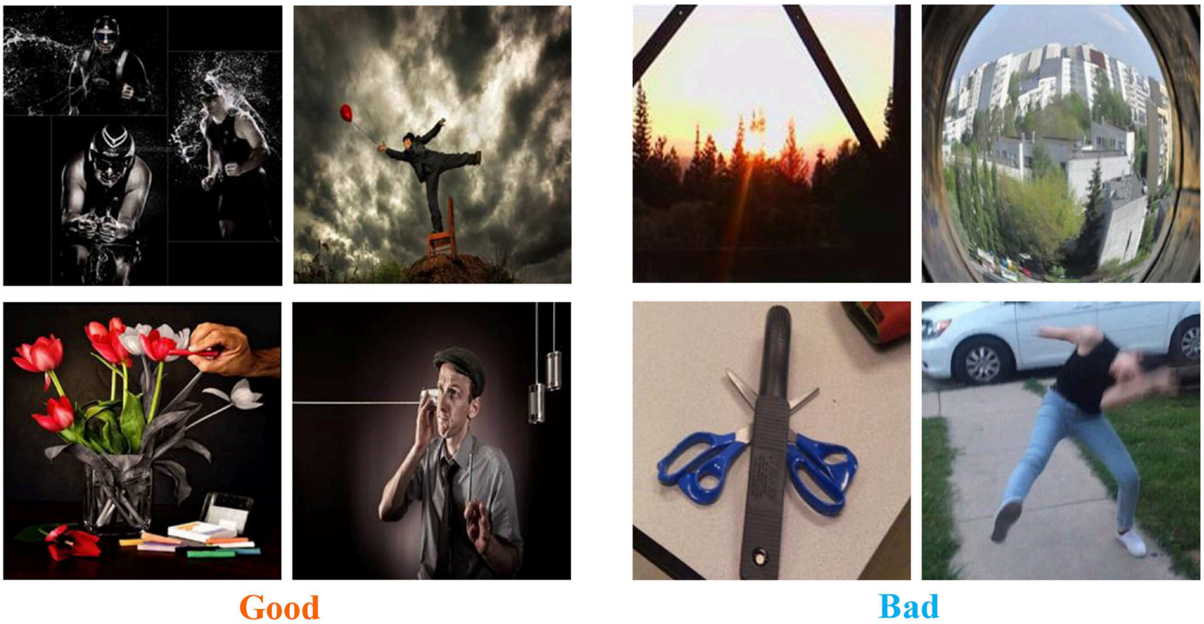


Fig. 3 *ILGNet architecture: Inception with connected local and global layers. We build this network on the first 1/3 part of GoogLeNetV1 [5] and batch normalisation, which is an important feature of GoogLeNetV2 [47]. One pre-treatment layer and three inception modules are used. We use the first two inception modules to compute the local features and the last one to compute global features. Connecting intermediate layers directly to the output layers has shown its value in recent work [5, 8]. Thus, we built a concat full-connected layer of 1024 dimension which connects two layers of local features and a layer of global features. The output layer indicates the probability of low- or high-aesthetic quality. The ILGNet contains 13 layers with parameters and without counting pooling layers (four layers). In Section 4, we use the labels (1)–(7) to demonstrate the visualisation results*

Table 2 Classification accuracy in AVA1 database

Methods	$\delta = 0$, %	$\delta = 1$, %
Traditional method [10]	66.70	67.00
RAPID [36]	69.91	71.26
RAPID-E [38]	74.46	73.70
Multi-patch [37]	75.41	—
AROD [53]	75.83	—
Multi-scene [40]	76.94	—
Comp.-prev. [11]	77.10	76.10
AADB [41]	77.33	—
BDN [43]	78.08	77.27
Semantic-based [42]	79.08	76.04
A-Lamp [44]	82.5	—
ILGNet-without-Inc.	**75.29**	**73.25**
1/3 GoogLeNetV1-BN	**80.74**	**79.09**
ILGNet-Inc.V1-BN	**81.68**	**80.71**
ILGNet-Inc.V3	**81.71**	**80.65**
ILGNet-Inc.V4	**82.66**	**80.83**

Table 3 Classification accuracy in AVA2 database

Methods	Accuracy, %
Subject-based [14]	61.49
EfficientAssess [49]	68.13
Generic-based [33]	68.55
Compt.-based [12]	68.67
High-level [13]	71.06
Multi-level [39]	78.92
Query-dependent [54]	80.38
DCNN-Aesth-SP [50]	83.52
Multi-scene [40]	84.88
ILGNet-without-Inc.	**79.64**
1/3 GoogLeNetV1-BN	**82.26**
ILGNet-Inc.V1-BN	**85.50**
ILGNet-Inc.V3	**85.51**
ILGNet-Inc.V4	**85.53**

Table 4 Efficiency comparison in AVA1 database

Methods	Accuracy $\delta = 0$, %	Training time, days	Test time, s
Full GoogLeNetV1-BN	**82.36**	16	0.84
2/3 GoogLeNetV1-BN	81.72	11	0.57
1/3 GoogLeNetV1-BN	80.74	**4**	0.33
ILGNet-Inc.V1-BN	81.68	**4**	**0.31**

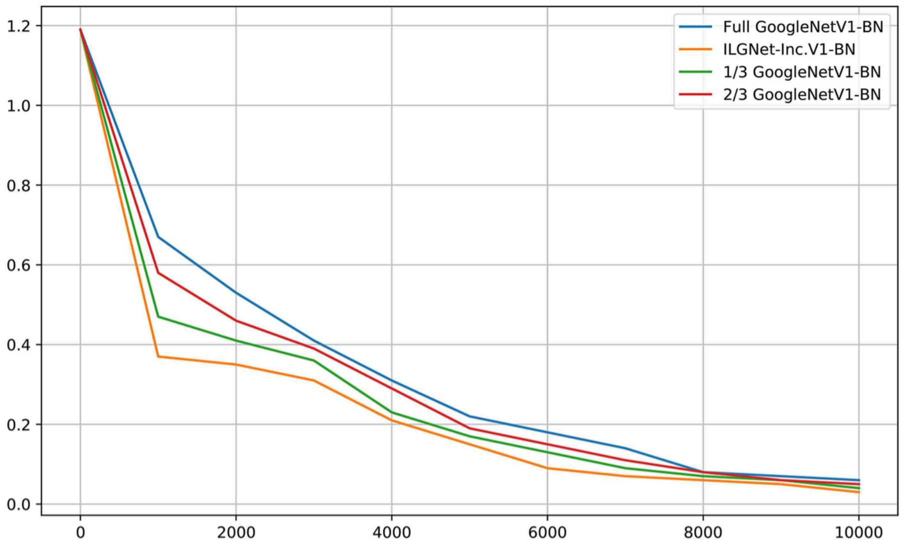
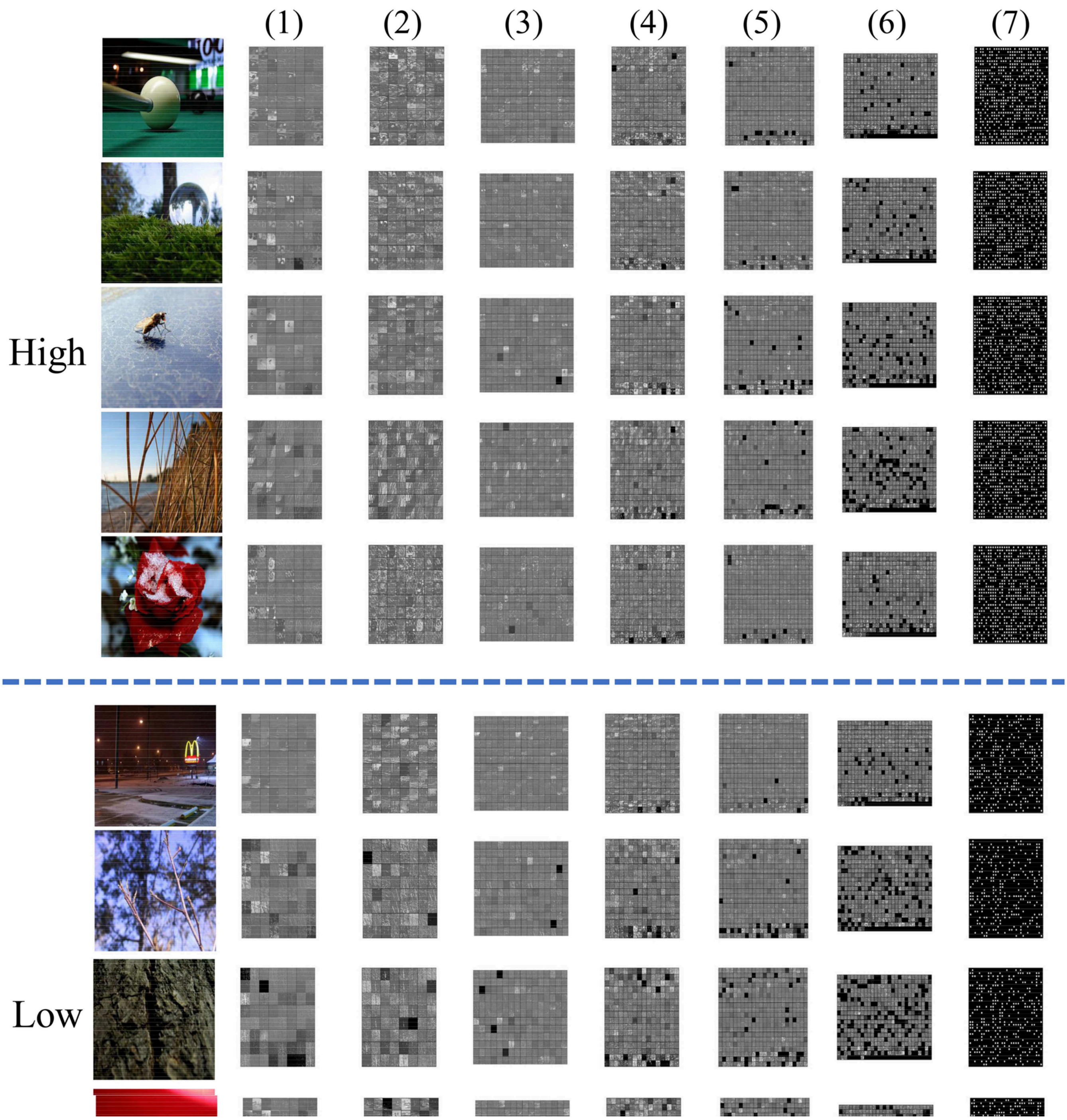


Fig. 4 *Loss versus epoch of our ILGNet-Inc.V1-BN, 1/3, 2/3 and full GoogLeNetV1-BN. in AVA1 database*


(1)
(2)
(3)
(4)
(5)
(6)
(7)
High
Low

[3] Li, J., Xia, C., Chen, X.: 'A benchmark dataset and saliency guided stacked autoencoders for video-based salient object detection', *IEEE Trans. Image Process.*, 2018, **27**, (1), pp. 349–364
[4] Jin, X., Wu, L., Li, X., *et al.*: 'Predicting aesthetic score distribution through cumulative Jensen–Shannon divergence', ArXiv, AAAI 2018, vol. abs/1708.07089, 2017
[5] Szegedy, C., Liu, W., Jia, Y., *et al.*: 'Going deeper with convolutions'. IEEE Conf. Computer Vision and Pattern Recognition CVPR, Boston, MA, USA, 7–12 June 2015, pp. 1–9
[6] Lin, M., Chen, Q., Yan, S.: 'Network in network', CoRR, vol. abs/1312.4400, 2013
[7] Arora, S., Bhaskara, A., Ge, R., *et al.*: 'Provable bounds for learning some deep representations'. Proc. 31st Int. Conf. Machine Learning ICML 2014, Beijing, China, 21–26 June 2014, pp. 584–592
[8] Maire, M., Yu, S.X., Perona, P.: 'Reconstructive sparse code transfer for contour detection and semantic labeling'. Computer Vision – ACCV – 12th Asian Conf. Computer Vision, Revised Selected Papers Part IV, Singapore, Singapore, 1–5 November 2014, pp. 273–287
[9] Jia, D., Wei, D., Richard, S., *et al.*: 'ImageNet: a large-scale hierarchical image database'. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009, pp. 248–255
[10] Murray, N., Marchesotti, L., Perronnin, F.: 'AVA: a large-scale database for aesthetic visual analysis'. IEEE Conf. Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012, pp. 2408–2415
[11] Mai, L., Jin, H., Liu, F.: 'Composition-preserving deep photo aesthetics assessment'. The IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 2016
[12] Datta, R., Joshi, D., Li, J., *et al.*: 'Studying aesthetics in photographic images using a computational approach'. Computer Vision – ECCV, Ninth European Conf. Computer Vision Proc. Part III, Graz, Austria, 7–13 May 2006, pp. 288–301
[13] Ke, Y., Tang, X., Jing, F.: 'The design of high-level features for photo quality assessment'. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR), New York, NY, USA, 17–22 June 2006, pp. 419–426
[14] Luo, Y., Tang, X.: 'Photo and video quality evaluation: focusing on the subject'. Computer Vision – ECCV 10th European Conf. Computer Vision Proc. Part III, Marseille, France, 12–18 October 2008, pp. 386–399
[15] Li, C., Chen, T.: 'Aesthetic visual quality assessment of paintings', *J. Sel. Top. Signal Process.*, 2009, **3**, (2), pp. 236–252
[16] Bhattacharya, S., Sukthankar, R., Shah, M.: 'A framework for photo-quality assessment and enhancement based on visual aesthetics'. Proc. 18th Int. Conf. Multimedia 2010, Firenze, Italy, 25–29 October 2010, pp. 271–280
[17] Jiang, W., Loui, A.C., Cerosaletti, C.D.: 'Automatic aesthetic value assessment in photographic images'. IEEE Int. Conf. Multimedia and Expo ICME 2010, Singapore, 19–23 July 2010, pp. 920–925
[18] Li, C., Gallagher, A.C., Loui, A.C., *et al.*: 'Aesthetic quality assessment of consumer photos with faces'. Proc. Int. Conf Image Processing ICIP 2010, Hong Kong, China, 26–29 September 2010, pp. 3221–3224
[19] Jin, X., Zhao, M., Chen, X., *et al.*: 'Learning artistic lighting template from portrait photographs'. Computer Vision – ECCV 2010 11th European Conf. Computer Vision Proc. Part IV, Heraklion, Crete, Greece, 5–11 September 2010, pp. 101–114
[20] Gray, D., Yu, K., Xu, W., *et al.*: 'Predicting facial beauty without landmarks'. Computer Vision – ECCV 2010–11th European Conf. Computer Vision Proc. Part VI, Heraklion, Crete, Greece, 5–11 September 2010, pp. 434–447
[21] Chen, X., Jin, X., Wu, H., *et al.*: 'Learning templates for artistic portrait lighting analysis', *IEEE Trans. Image Process.*, 2015, **24**, (2), pp. 608–618
[22] Dhar, S., Ordonez, V., Berg, T.L.: 'High level describable attributes for predicting aesthetics and interestingness'. The 24th IEEE Conf. Computer Vision and Pattern Recognition CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011, pp. 1657–1664
[23] Joshi, D., Datta, R., Fedorovskaya, E.A., *et al.*: 'Aesthetics and emotions in images', *IEEE Signal Process. Mag.*, 2011, **28**, (5), pp. 94–115
[24] Nishiyama, M., Okabe, T., Sato, I., *et al.*: 'Aesthetic quality classification of photographs based on color harmony'. The 24th IEEE Conf. Computer Vision Computer Vision ICCV 2011, Barcelona, Spain, 6–13 November 2011, pp. 1784–1791
[34] Lin, J., He, C., Jane Wang, Z., *et al.*: 'Structure preserving transfer learning for unsupervised hyperspectral image classification', *IEEE Geosci. Remote Sens. Lett.*, 2017, **14**, pp. 1656–1660
[35] Karayev, S., Trentacoste, M., Han, H., *et al.*: 'Recognizing image style'. British Machine Vision Conf. BMVC 2014, Nottingham, UK, 1–5 September 2014
[36] Lu, X., Lin, Z., Jin, H., *et al.*: 'RAPID: rating pictorial aesthetics using deep learning'. Proc. ACM Int. Conf. Multimedia MM'14, Orlando, FL, USA, 03–07 November 2014, pp. 457–466
[37] Lu, X., Lin, Z., Shen, X., *et al.*: 'Deep multi-patch aggregation network for image style, aesthetics, and quality estimation'. 2015 IEEE Int. Conf. Computer Vision ICCV 2015, Santiago, Chile, 7–13 December 2015, pp. 990–998
[38] Lu, X., Lin, Z.L., Jin, H., *et al.*: 'Rating image aesthetics using deep learning', *IEEE Trans. Multimed.*, 2015, **17**, (11), pp. 2021–2034
[39] Dong, Z., Tian, X.: 'Multi-level photo quality assessment with multi-view features', *Neurocomputing*, 2015, **168**, pp. 308–319
[40] Wang, W., Zhao, M., Wang, L., *et al.*: 'A multi-scene deep learning model for image aesthetic evaluation', *Signal Process. Image Commun.*, 2016, **47**, pp. 511–518
[41] Kong, S., Shen, X., Lin, Z., *et al.*: 'Photo aesthetics ranking network with attributes and content adaptation'. European Conf. Computer Vision (ECCV), Amsterdam, Netherlands, 2016
[42] Kao, Y., He, R., Huang, K.: 'Deep aesthetic quality assessment with semantic information', *IEEE Trans. Image Process.*, 2017, **26**, (3), pp. 1482–1495
[43] Wang, Z., Liu, D., Chang, S., *et al.*: 'Image aesthetics assessment using Deep Chatterjee's machine'. Int. Joint Conf. Neural Networks (IJCNN), Anchorage, AK, USA, 2017
[44] Ma, S., Liu, J., Chen, C.W.: 'A-Lamp: adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment'. 2017 IEEE Conf. Computer Vision and Pattern Recognition CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 722–731
[45] Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'ImageNet classification with deep convolutional neural networks'. Advances in Neural Information Processing Systems 25: 26th Annual Conf. Neural Information Processing Systems 2012 Proc. meeting held 3–6 December 2012, Lake Tahoe, NV, USA, 2012, pp. 1106–1114
[46] Simonyan, K., Zisserman, A.: 'Very deep convolutional networks for large-scale image recognition', CoRR, vol. abs/1409.1556, 2014
[47] Ioffe, S., Szegedy, C.: 'Batch normalization: accelerating deep network training by reducing internal covariate shift'. Proc. 32nd Int. Conf. Machine Learning ICML 2015, Lille, France, 6–11 July 2015, pp. 448–456
[48] Jia, Y., Shelhamer, E., Donahue, J., *et al.*: 'Caffe: convolutional architecture for fast feature embedding', arXiv preprint arXiv:1408.5093, 2014
[49] Lo, K.-Y., Liu, K.-H., Chen, C.-S.: 'Assessment of photo aesthetics with efficiency'. Proc. 21st Int. Conf. Pattern Recognition ICPR 2012, Tsukuba, Japan, 11–15 November 2012, pp. 2186–2189
[50] Dong, Z., Shen, X., Li, H., *et al.*: 'Photo quality assessment with DCNN that understands image well'. MultiMedia Modeling – 21st Int. Conf., MMM 2015 Proc. Part II, Sydney, NSW, Australia, 5–7 January 2015, pp. 524–535
[51] Szegedy, C., Vanhoucke, V., Ioffe, S., *et al.*: 'Rethinking the inception architecture for computer vision'. 2016 IEEE Conf. Computer Vision and Pattern Recognition CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 2818–2826
[52] Szegedy, C., Ioffe, S., Vanhoucke, V., *et al.*: 'Inception-V4, Inception-ResNet and the impact of residual connections on learning'. Proc. 31st AAAI Conf. Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017, pp. 4278–4284
[53] Schwarz, K., Wieschollek, P., Lensch, H.P.A.: 'Will people like your image?', CoRR, vol. abs/1611.05203, 2016
[54] Tian, X., Dong, Z., Yang, K., *et al.*: 'Query-dependent aesthetic model with deep learning for photo quality assessment', *IEEE Trans. Multimed.*, 2015, **17**, (11), pp. 2035–2048