Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11295. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05710-7_30

Image Aesthetics Assessment using Fully Convolutional Neural Networks

Konstantinos Apostolidis and Vasileios Mezaris

Information Technologies Institute / CERTH
6th km Charilaou - Thermi Road, Thermi - Thessaloniki
{kapost,bmezaris}@iti.gr

Abstract. This paper presents a new method for assessing the aesthetic quality of images. Based on the findings of previous works on this topic, we propose a method that addresses the shortcomings of existing ones, by: a) Making possible to feed higher-resolution images in the network, by introducing a fully convolutional neural network as the classifier. b) Maintaining the original aspect ratio of images in the input of the network, to avoid distortions caused by re-scaling. And c) combining local and global features from the image for making the assessment of its aesthetic quality. The proposed method is shown to achieve state of the art results on a standard large-scale benchmark dataset.

Keywords: Image Aesthetics, Deep Learning, Fully Convolutional Neural Networks

1 Introduction

Aesthetic quality assessment is an established task in the field of image processing and aims at computationally distinguishing high aesthetic quality photos from low aesthetic quality ones. Aesthetic quality assessment solutions can contribute to applications and tasks such as image re-ranking [35, 31], search and retrieval of photos [27] and videos [14], image enhancement methods [1, 9] and image collection summarization and preservation [31, 26]. The automatic prediction of a photo's aesthetic value is a challenging problem because, among others, humans often assess the aesthetic quality based on their subjective criteria; thus, it is difficult to define a clear and subjective set of rules for automating this assessment.

In this paper, we present an automatic aesthetic assessment method based on a fully convolutional neural network that utilizes skip connections and a setup for minimizing the sizing distortions of the input image. The rest of the paper is organized as follows: in Section 2 we review the related work. In Section 3 we present the proposed method in detail. This is followed by reporting the experimental setup, results and comparisons in Section 4, and finally we draw conclusions and provide a brief future outlook in Section 5.

2 Related Work

The early attempts on image aesthetic quality assessment used handcrafted features, such as the methods of [24] and [22]. Both of these methods base their features on photographic rules that usually apply in aesthetically appealing photos. The method of [17] also uses handcrafted features but with a focus on efficiency.

Due to the success of deep convolutional neural networks (DCNN) on image classification [30, 32] and transfer learning [5], more recent attempts are based on the use of DCNNs. To our knowledge, the first of such methods is [19]. In [19] a deep learning system is introduced (RAPID - RAting PIctorial aesthetics using Deep learning) that aims to incorporate heterogeneous inputs generated from the image, which include a global view and local views. The global view is represented by a normalized-to-square-size input, while local views are represented by small randomly-cropped square parts of the original high-resolution image. Additionally, the method of [19] utilizes certain style attributes of images (e.g. "color harmony", "good lighting", "object emphasis", "vivid color", etc.) to help improve the aesthetic quality categorization accuracy; however, generating these attribute annotations may result in high inference times. In a later work [20], the same authors employ the style and semantic attributes of images to further boost the aesthetic categorization performance. [21] claims that the constraint of the neural networks to take a fixed- and squared-size image as input (i.e. images need to be transformed via cropping, scaling, or padding) compromises the assessment of the aesthetic quality of the original images. To alleviate this, [21] presents a composition-preserving deep convolutional network method that directly learns aesthetic features from the original input images without any image transformations.

In [12] its authors argue that the two classes of high and low aesthetic qualities contain large intra-class differences, and propose a model to jointly learn meaningful photographic attributes and image content information that can help regularize the complicated photo aesthetic rating problem. To train their model, they assemble a new aesthetics and attributes database (AADB).

In [2] its authors investigate the use of a DCNN to predict image aesthetics by fine-tuning a canonical CNN architecture, originally trained to classify objects and scenes, casting the image aesthetic quality prediction as a regression problem. They also investigate whether image aesthetic quality is a global or local attribute, and the role played by bottom-up and top-down salient regions to the prediction of the global image aesthetics. In [11], its authors aiming once again to take both local and global features of images into consideration, propose a DCNN architecture named ILGNet, which combines both the Inception modules and a connected layer of both local and global features. 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, resulting in a 1024-dimension layer. Finally, in [7], a complex framework for aesthetic quality assessment is introduced. Specifically, the authors design several rule-based aesthetic features, and also use content-based features extracted with the help of a DCNN. They claim that these two type of features are complementary to each other, and combine them using a Multi Kernel Learning method. To our knowledge, this method achieves the state of the art results on the popular AVA2 dataset.

Finally, we should note that there are several works that deal with the relation between users' preferences and the assessment of the aesthetic quality of photos, such as [3, 4, 29, 34]. However, this is out of the scope of our present work, since we are addressing the problem of user-independent prediction of image aesthetic quality similarly to [11, 2, 12, 21, 7, 33, 24] and many other works.

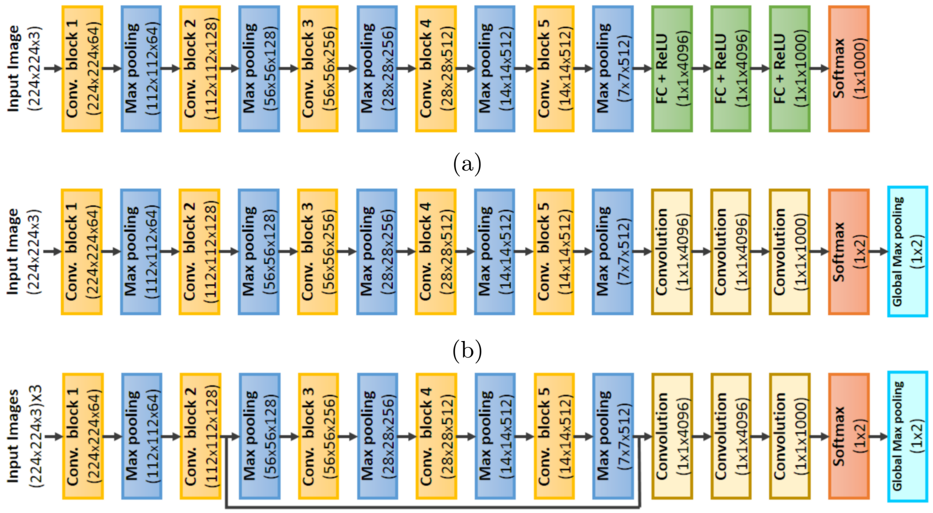
From the review of the related work, it can be easily asserted that after the introduction of DCNNs for aesthetic quality assessment the main effort has focused on two directions: a) minimizing the sizing distortions of the input image; b) combining local and global features to facilitate the aesthetics assessment. Inspired by there, we set three objectives: a) using a fully convolutional neural network, to experiment with feeding higher-resolution images to the network (this is done in a way that weights can be copied from a pre-trained model, without needing to re-train the network from scratch); b) introducing an approach for maintaining the aspect ratio of the input image; c) introducing a skip connection in our network to combine the output from early layers to that of the later layers, thus introducing information from local features to the final decision of the network.

3 Proposed Method

A fully connected (FC) layer has nodes connected to all activations in the previous layer, hence, requires a fixed size of input data. It is worth noting that the only difference between an FC layer and a convolutional layer is that the neurons in the convolutional layer are connected only to a local region in the input. However, the neurons in both layers still compute dot products, so their functional form is identical.

Therefore, our first step is to convert the network to a fully convolutional network (FCN). To do so we must change the FC layers to convolutional layers (see Fig. 1a and 1b). For the purpose of this paper we use the VGG16 architecture [30] for simplicity - yet our method can be applied to any DCNN architecture with little modification. This architecture has three FC layers at the end of the network. We can convert each of these three FC layers to convolutional layers as follows:

- Replace the first FC layer that requires a 7×7×512 tensor with a convolutional layer that uses filter size equal to 7, giving an output tensor of 1×1×4096 dimension.
- Replace the second FC layer with a CONV layer that uses filter size equal to 1, giving an output tensor of 1×1×4096 dimension;
- Replace the last FC layer similarly, with filter size equal to 1, giving the final output tensor of 1×1×2 dimension since we want to fine-tune the network for the two-class aesthetic quality assessment problem.



(c)

Fig. 1: Network models used in this work: a) the original VGG16 network, trained for 1000 ImageNet classes, b) the fully convolutional version of VGG16, for 2 classes (high aesthetic quality, low aesthetic quality), c) the proposed fully convolutional VGG16, with an added skip connection (after the second convolutional block to the decision convolutional layers) and accepting a triplet of image croppings as input. In all the above model illustrations the following color-coding is used: yellow for convolutional layers, dark yellow for blocks of convolutional layers, green for fully connected layers, orange for softmax operations, blue for max pooling operations, light blue for global max pooling operations.

This conversion allows us to "slide" the original convolutional network very efficiently across many spatial positions in a larger image, in a single forward pass, an advantage which is known in the literature; FCNs were first used in [23] to classify series of handwritten digits and more recently for semantic segmentation [18]. Additionally, each of these conversions could in practice involve manipulating (i.e. reshaping) the weight matrix in each FC layer into the weights of the convolutional layer filters. Therefore, we can easily copy the weights of a pre-trained VGG16 on ImageNet [13]. This in turn, allows for faster training times and does not require a large collection of training images, since the network is not trained from scratch.

One thing to note here is that since we "slide" the convolutional network in the image, the FCN produces many decisions, one for each spatial region analyzed. Therefore, to come up with a single decision and to be able to re-train the network we add on top of the FCN a global pooling operation layer for spatial data. This can be either a global max pooling layer or a global average pooling layer. In the experiments conducted in Section 4, we test both approaches.

Regarding our objective to maintain the original aspect ratio, there are various known approaches: a) cropping the center part of the image (and discarding the cropped parts), b) padding the image (adding blank borders) to make it of square size, c) feeding the image in an FCN at its original size, d) feeding multiple croppings of the image to ensure that the whole surface is scanned by the network (even though overlapping of the scanned regions may occur). The third of the above approaches can only be achieved if an FCN is utilized. In the case of the second option (padding) some literature works argue that introducing blank parts in the input image can greatly deteriorate the performance of the network. Thus, we examine one more variation in which the input image is fed as a padded and masked square image. To achieve this, we input to the network a binary mask (containing ones for the areas that exist in the original image and zeros for the added black areas). An element-wise multiplication takes place before the decision layers (namely, the convolutional layers that replaced the FC layers of the original model) to zero the filters output in the blank areas of the image. Another approach, in the spirit of performing multiple croppings (but not previously used for aesthetics assessment), is proposed in the present work. As shown in Fig. 2, three overlapping croppings of each input image are jointly fed into the network. All of the aforementioned approaches are evaluated in Section 4.

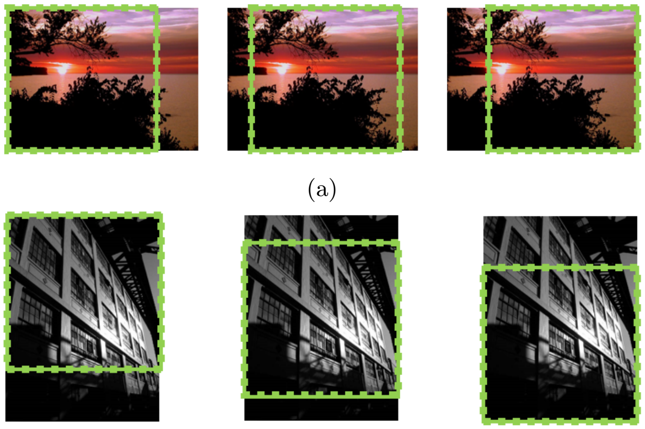
The notion of introducing skip connections in a neural network is known in the literature (in different application domains, such as biomedical image segmentation [8]). We should note here that this is different to connecting multiple layers in a network as in [16], or the way used in the Dense architecture of neural networks [10]: skip connections aim to combine the output from a single early layer with the decision made in the last layers. However, the choice of which early layer's output to use is not an easy one; the results of extensive experiments regarding the effect of using skip connections in DCNNs on classifying images in [15] show that this choice heavily depends on the specific application domain. Tests were reported in [15] on seven datasets of different nature (classification of gender, texture, recognition of digits and objects). Since the aesthetic quality assessment problem is probably more closely related to texture classification (compared to the other application domains examined in [15]) and based upon the observation reported in [15], we choose to introduce a skip connection from immediately after the second convolution block to the layer prior to decision layers (i.e. the convolutional layers that replaced the FC layers of the original model, see Fig. 1c).

4 Experimental Results

4.1 Dataset

The Aesthetic Visual Analysis (AVA) dataset [25] is a list of image ids from DPChallenge.com, which is a on-line photography social network. There are in total 255529 photos, each of which is rated by a large number of persons. The range of the scores used for the rating is 1-10. We choose to use the AVA dataset

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11295. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05710-7_30



(a)

(b)

Fig. 2: Illustration of the proposed three croppings with respect to the original image a) for images in landscape mode, b) for images in portrait mode.

since it is the largest in the domain of aesthetic quality assessment. Two, widely used, ways of splitting the AVA dataset into training and test portions are found in the literature:

- AVA1: The score of 5 is chosen as the threshold to distinguish the AVA images to high and low aesthetic quality. This way 74673 images are labeled as of high aesthetic quality and 180856 are labeled as of low aesthetic quality. The dataset is randomly split into the training set (totaling 234599 images) and testing set (19930 images) [25, 33, 34, 12].
- AVA2: The images in the AVA dataset are sorted according to their mean aesthetic quality score. The top 10% images are labeled as of good aesthetic quality and the bottom 10% are labeled as of bad aesthetic quality. This way 51106 images are used from the dataset. These images are randomly divided into 2 equally-sized sets, which are the training and testing sets, respectively [12, 7, 25, 6, 33, 24, 14].

Similarly to most of the recent literature works, we choose to use the AVA2 dataset in the experiments conducted in the sequel, since the way that it is constructed ensures the reliability of the ground-truth aesthetic quality annotations.

4.2 Experimental Setup

As already mentioned we base our proposed FCN on the VGG16 [30] architecture for the sake of simplicity, yet our method can be applied to any DCNN architecture with limited modifications. For implementation, we used the Keras neural network API[1]. Our experimental setup regarding the tests conducted in

[1] https://keras.io/

this section is as follows: we set the starting learning rate to 0.01; and used a callback function of Keras to reduce the learning rate if the validation accuracy is not increased for three consecutive epochs. The batch size was fixed to 8, unless noted otherwise. We set the number of training epochs to 40. The results reported here are the achieved accuracy by using the model after the 40th epoch. The code for converting VGG16 (as well as numerous other architectures) to an FCN, the implementation of skip connections and the methods tested for maintaining the original aspect ratio are made publicly available online[2]. All experiments were conducted on a PC with an i7-4770K CPU, 16GB of RAM and a Nvidia GTX 1080 Ti GPU.

4.3 Results

We first conducted some preliminary experiments in order to test: a) the relation of input image size to the accuracy. The performance of all tested setups was evaluated both in terms of detection accuracy and the time-efficiency (measuring the average inference time for a single image). Table 1 reports the results of each compared approach. The first column of this table cites the name of the used network. The second column reports the input image size. We performed experiments by resizing the AVA2 images to: a) the size that the VGG16 model was originally trained for (224×224), b) 1.5× this original VGG16 size, c) 2× the original VGG16 size and d) 3× the original VGG16 size, resulting to testing images finally resized to size 224×224, 336×336, 448×448 and 672×672 pixels, respectively. We also performed experiments where we fed the input image resizing its height to 336 pixels and accordingly adjusting its width in order to maintain its original aspect ratio (denoted as "336×A.W" in the last four rows of Table 1). In the third column, we report the batch size used during the training phase. As already mentioned this was fixed to the value of 8 except for the experiments in the last four rows of Table 1, since in these specific setups images of different sizes cannot be fed into the network in a single batch. The fourth column reports whether we freeze any layers (i.e. not updating the weights of these layers) or not. The fifth column reports the type of global pooling applied at the end of the network (only when using the proposed FCN; not applicable when using the original VGG16). Finally in the last two columns we report the average inference time and the accuracy achieved in the AVA2 dataset. Examining Table 1, we observe that increasing the input image size does not necessarily improve the results. Specifically, increasing the input size from 224×224 to 336×336 achieved better accuracy in all cases. However, further increasing the input size from 336×336 to 448×448 or 672×672 consistently led to slight reduction of the performance of the network. In the cases where we adjusted the images' height to the fixed value of 336 pixels while maintaining the original aspect ratio, the network yielded very poor performance, mainly due to using a batch size equal to 1. Additionally, with respect to the time-efficiency, we observe that increasing

[2] Implementation of fully convolutional networks in Keras is available at https://github.com/bmezaris/fully_convolutional_networks

Table 1: Results of preliminary tests.

Setup used	Input size (h.×w.)	Batch size	Freeze	Global Pooling	Infer time (avg ± dev.) (ms)	AVA2 accuracy (%)
VGG16	224×224	8	Yes	N/A	110 ± 5	84.03
VGG16	224×224	8	No	N/A	110 ± 5	85.04
FCN	224×224	8	Yes	Max	120 ± 5	84.57
FCN	224×224	8	Yes	Average	120 ± 5	84.96
FCN	224×224	8	No	Max	120 ± 5	86.20
FCN	224×224	8	No	Average	120 ± 5	85.06
FCN	336×336	8	Yes	Max	160 ± 5	88.35
FCN	336×336	8	Yes	Average	160 ± 5	88.26
FCN	336×336	8	No	Max	160 ± 5	**88.44**
FCN	336×336	8	No	Average	160 ± 5	88.21
FCN	448×448	8	Yes	Max	480 ± 5	87.65
FCN	448×448	8	Yes	Average	480 ± 5	87.35
FCN	448×448	8	No	Max	480 ± 5	88.01
FCN	448×448	8	No	Average	480 ± 5	86.91
FCN	672×672	8	Yes	Max	790 ± 5	86.03
FCN	672×672	8	Yes	Average	790 ± 5	85.66
FCN	672×672	8	No	Max	790 ± 5	87.52
FCN	672×672	8	No	Average	790 ± 5	87.07
FCN	336×A.W.	1	Yes	Max	280 ± 100	66.02
FCN	336×A.W.	1	Yes	Average	280 ± 100	61.28
FCN	336×A.W.	1	No	Max	280 ± 100	73.02
FCN	336×A.W.	1	No	Average	280 ± 100	71.17

the image size quadratically increases the inference time for a single image. The average inference time of 790 ms for the 672×672 size of input image is possibly prohibitively high for real-world applications, which is an additional reason to not use such large input sizes.

Regarding the freezing of layers during the fine-tuning process we tested two approaches: a) freezing the first layers up to the end of the second convolutional block of VGG16 (denoted as "Yes" in the fourth column of Table 1), and b) not freezing any layer (denoted as "No" in the fourth column of Table 1). It is known in the literature [28] that the weights of the first network layers can remain frozen, i.e., they are copied from the pre-trained DCNN and kept unchanged, since these learn low-level image characteristics which are useful for most types of image classification. However, as can be asserted from Table 1, not freezing any layer consistently gives better accuracy. This can be explained from the fact that the problem of aesthetic quality assessment is quite different from image classification in ImageNet. Thus, it is better to let the network adjust the weights of all its layers.

Concerning the type of global pooling applied at the end of the network, we notice that using global max pooling in most cases yields better results.

Therefore, for the next set of experiments: a) we use the global max pooling operation as the last layer in the network, b) we do not freeze any layer during the fine-tuning process, and c) we input images of size 336×336 to the network.

Table 2: Results of tests regarding methods preserving the aspect ratio of the original images.

Setup used	Infer time (avg ± dev.) (ms)	AVA2 accuracy (%)
FCN	160 ± 5	88.44
FCN + padding	110 ± 5	86.08
FCN + cropping	110 ± 5	86.53
FCN + masking	120 ± 5	87.61
FCN + 3× croppings	150 ± 5	**89.94**

We proceed to conduct experiments to test different approaches for maintaining the original aspect ratio on the best performing setup of Table 1. The results are reported in Table 2 and the result of the best performing setup of Table 1 is copied in the first row of the new table. We notice that the first three approaches reported in Table 2 (“FCN + padding”, “FCN + cropping”, “FCN + masking”) lead to lower the accuracy, compared to not maintaining the original aspect ratio (i.e. resizing images to 336×336 pixels). Contrary to this, the proposed last approach of Table 2 that uses three croppings of the original image to include all the surface of the image in the network exhibits increased accuracy, reaching 89.94%.

Table 3: Results of tests regarding the effect of adding a skip connection to the network.

Setup used	Infer time (avg ± dev.) (ms)	AVA2 accuracy (%)
FCN + masking	120 ± 5	87.61
FCN + 3× croppings	150 ± 5	89.74
FCN + masking + skip connection	120 ± 5	83.40
FCN + 3× croppings + skip connection	150 ± 5	**91.01**

Then we test the effect of adding a skip connection to the best performing setup of Table 2. The new results are reported in Table 3 and the results of the “FCN + masking” and “FCN + 3× croppings” setups from Table 2 are copied in the first two rows of the new table. We observe that introducing a skip connection improves the achieved accuracy in the case of “FCN + 3× croppings” setup. On the other hand, introducing the skip connection on the "masking" setup considerably reduces the accuracy, since the values of the filters that where excluded using the mask are re-introduced in the decision layer.

Finally, the "FCN + 3× croppings + skip connection", which is the method proposed in this work, is shown in Table 2 to achieve state of the art results, outperforming [24, 33, 11, 7] that report accuracy scores of up to 90.76% on the AVA2 dataset. This is achieved even though the VGG16 architecture, that our network is based on, is not the most powerful deep network architecture, as documented by the literature on object/image annotation and other similar problems.

Table 4: Comparison of the proposed method to methods of the literature.

Method	AVA2 accuracy (%)
Handcrafted features[24]	77.08
MSDLM [33]	84.88
ILGNet [11]	85.62
MKL_3 [7]	90.76
Proposed (FCN + 3× croppings + skip connection)	**91.01**

5 Conclusions

In this paper we presented a method for assessing the aesthetic quality of images. Drawing inspiration from the related literature we converted a deep convolutional neural network to a fully convolutional network, in order to be able to feed images of arbitrary size to the network. A variety of conducted experiments provided useful insight regarding the tuning of parameters of our proposed network. Additionally, we proposed an approach for maintaining the original aspect ratio of the input images. Finally, we introduced a skip connection in the network, to combine local and global information of the input image in the aesthetic quality assessment decision. Combining all the proposed techniques we achieve state of the art results as can be ascertained by our experiments and comparisons. In the future, we plan to examine the impact of these proposed techniques on different network architectures.

6 Acknowledgments

This work was supported by the EUs Horizon 2020 research and innovation programme under contracts H2020-687786 InVID and H2020-732665 EMMA.

References

1. Bhattacharya, S., Sukthankar, R., Shah, M.: A framework for photo-quality assessment and enhancement based on visual aesthetics. In: Proc. 18th ACM Int. Conf. on Multimedia (MM). pp. 271–280. ACM (2010)
2. Bianco, S., Celona, L., Napoletano, P., Schettini, R.: Predicting image aesthetics with deep learning. In: Proc. Int. Conf. on Advanced Concepts for Intelligent Vision Systems (ACIVS). pp. 117–125. Springer (2016)
3. Cui, C., Fang, H., Deng, X., Nie, X., Dai, H., Yin, Y.: Distribution-oriented aesthetics assessment for image search. In: Proc. 40th Int. SIGIR Conf. on Research and Development in Information Retrieval. pp. 1013–1016. ACM (2017)
4. Deng, X., Cui, C., Fang, H., Nie, X., Yin, Y.: Personalized image aesthetics assessment. In: Proc. Conf. on Information and Knowledge Management. pp. 2043–2046. ACM (2017)
5. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: Proc. Int. Conf. on Machine Learning (ICML). pp. 647–655 (2014)
6. Dong, Z., Shen, X., Li, H., Tian, X.: Photo quality assessment with DCNN that understands image well. In: Proc. Int. Conf. on Multimedia Modeling (MMM). pp. 524–535. Springer (2015)
7. Dong, Z., Tian, X.: Multi-level photo quality assessment with multi-view features. Neurocomputing **168**, 308–319 (2015)
8. Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C.: The importance of skip connections in biomedical image segmentation. In: Deep Learning and Data Labeling for Medical Applications, pp. 179–187. Springer (2016)
9. Guo, Y., Liu, M., Gu, T., Wang, W.: Improving photo composition elegantly: Considering image similarity during composition optimization. In: Computer graphics forum. vol. 31, pp. 2193–2202. Wiley Online Library (2012)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proc. Conf on Computer Vision and Pattern Recognition (CVPR). vol. 1, p. 3 (2017)
11. Jin, X., Chi, J., Peng, S., Tian, Y., Ye, C., Li, X.: Deep image aesthetics classification using inception modules and fine-tuning connected layer. In: Proc. IEEE 8th Int. Conf. on Wireless Communications & Signal Processing (WCSP). pp. 1–6. IEEE (2016)
12. Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking network with attributes and content adaptation. In: Proc. European Conf. on Computer Vision (ECCV). pp. 662–679. Springer (2016)
13. Krizhevsky, A., Ilya, S., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Proc. Conf. on Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105. Curran Associates, Inc. (2012)
14. Lemarchand, F.: From computational aesthetic prediction for images to films and online videos. Avant **8**, 69–78 (2017)
15. Li, Y., Zhang, T., Liu, Z., Hu, H.: A concatenating framework of shortcut convolutional neural networks. arXiv preprint arXiv:1710.00974 (2017)
16. Liang, M., Hu, X., Zhang, B.: Convolutional neural networks with intra-layer recurrent connections for scene labeling. In: Proc. Conf. on Advances in Neural Information Processing Systems (NIPS). pp. 937–945. Red Hook, NY Curran (2015)
17. Lo, K.Y., Liu, K.H., Chen, C.S.: Assessment of photo aesthetics with efficiency. In: Proc. 21st Int. Conf. on Pattern Recognition (ICPR). pp. 2186–2189. IEEE (2012)

18. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proc. Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 3431–3440. IEEE (2015)
19. Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.Z.: Rapid: Rating pictorial aesthetics using deep learning. In: Proc. 22nd ACM Int. Conf. on Multimedia (MM). pp. 457–466. ACM (2014)
20. Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.Z.: Rating image aesthetics using deep learning. IEEE Trans. on Multimedia **17**(11), 2021–2034 (2015)
21. Mai, L., Jin, H., Liu, F.: Composition-preserving deep photo aesthetics assessment. In: Proc. Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 497–506. IEEE (2016)
22. Marchesotti, L., Perronnin, F., Larlus, D., Csurka, G.: Assessing the aesthetic quality of photographs using generic image descriptors. In: Proc. IEEE Int. Conf. on Computer Vision (ICCV). pp. 1784–1791. IEEE (2011)
23. Matan, O., Burges, C.J., LeCun, Y., Denker, J.S.: Multi-digit recognition using a space displacement neural network. In: Proc. Conf. on Advances in Neural Information Processing Systems (NIPS). pp. 488–495 (1992)
24. Mavridaki, E., Mezaris, V.: A comprehensive aesthetic quality assessment method for natural images using basic rules of photography. In: Proc. IEEE Int. Conf. on Image Processing (ICIP). pp. 887–891. IEEE (2015)
25. Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: Proc. Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 2408–2415. IEEE (2012)
26. Nejdl, W., Niederee, C.: Photos to remember, photos to forget. IEEE Trans. on MultiMedia (TMM) **22**(1), 6–11 (2015)
27. Obrador, P., Anguera, X., de Oliveira, R., Oliver, N.: The role of tags and image aesthetics in social image search. In: Proc. 1st SIGMM workshop on Social media. pp. 65–72. ACM (2009)
28. Pittaras, N., Markatopoulou, F., Mezaris, V., Patras, I.: Comparison of fine-tuning and extension strategies for deep convolutional neural networks. In: Proc. 23rd Int. Conf. on Multimedia Modeling (MMM). pp. 102–114. Springer (2017)
29. Ren, J., Shen, X., Lin, Z.L., Mech, R., Foran, D.J.: Personalized image aesthetics. In: Proc. IEEE Int. Conf. on Computer Vision (ICCV). pp. 638–647. IEEE (2017)
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proc. Int. Conf. on Learning Representations (ICLR) (2015)
31. Su, H.H., Chen, T.W., Kao, C.C., Hsu, W.H., Chien, S.Y.: Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features. IEEE Trans. on Multimedia (TMM) **14**(3), 833–843 (2012)
32. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proc. Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 1–9. IEEE (2015)
33. Wang, W., Zhao, M., Wang, L., Huang, J., Cai, C., Xu, X.: A multi-scene deep learning model for image aesthetic evaluation. Signal Processing: Image Communication **47**, 511–518 (2016)
34. Wang, Z., Liu, D., Chang, S., Dolcos, F., Beck, D., Huang, T.: Image aesthetics assessment using deep chatterjee's machine. In: Proc. IEEE Int. Joint Conf. on Neural Networks (IJCNN). pp. 941–948. IEEE (2017)
35. Yeh, C.H., Ho, Y.C., Barsky, B.A., Ouhyoung, M.: Personalized photograph ranking and selection system. In: Proc. 18th ACM Int. Conf. on Multimedia (MM). pp. 211–220. ACM (2010)