EasyChair Preprint

№ 4112

A Deep Drift-Diffusion Model for Image Aesthetic Score Distribution Prediction

Xin Jin, Xiqiao Li, Heng Huang, Xiaodong Li and Xinghui Zhou

August 30, 2020

A Deep Drift-Diffusion Model for Image Aesthetic Score Distribution Prediction

Xin Jin[a] Xiqiao Li [a] Heng Huang [a] Xiaodong Li [a]and Xinghui Zhou[a]

[a]Beijing Electronic Science and Technology Institute;

ABSTRACT

The task of aesthetic quality assessment is complicated due to its subjectivity. In recent years, the target representation of image aesthetic quality has changed from a one-dimensional binary classification label or numerical score to a multi-dimensional score distribution. According to current methods, the ground truth score distributions are straightforwardly regressed. However, the subjectivity of aesthetics is not taken into account, that is to say, the psychological processes of human beings are not taken into consideration, which limits the performance of the task. In this paper, we propose a Deep Drift-Diffusion (DDD) model inspired by psychologists to predict aesthetic score distribution from images. The DDD model can describe the psychological process of aesthetic perception instead of traditional modeling of the results of assessment. We use deep convolution neural networks to regress the parameters of the drift-diffusion model. The experimental results in large scale aesthetic image datasets reveal that our novel DDD model is simple but efficient, which outperforms the state-of-the-art methods in aesthetic score distribution prediction. Besides, different psychological processes can also be predicted by our model.

Keywords: neural networks, aesthetic assessment, distribution prediction, psychology process

1. INTRODUCTION

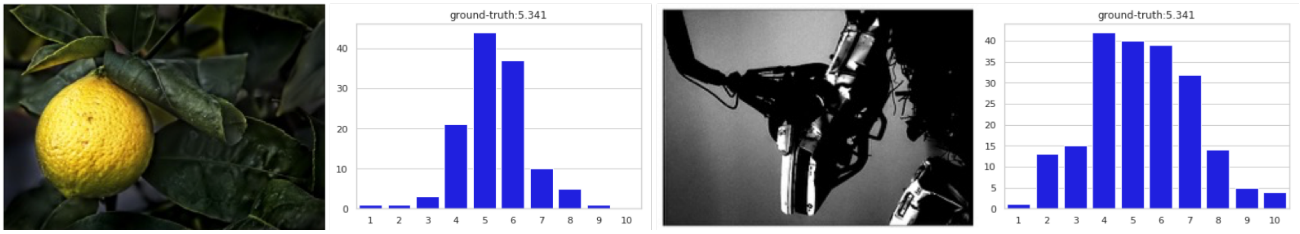

Figure 1. Examples images and their aesthetic score distribution from AVA dataset. The aesthetic scores are rated from 1 to 10 and the vertical axis of the histogram represents the number of ratings.

Given the explosive growth of digital photography in the Internet and social networks, image aesthetic evaluation has received more and more attention due to its huge application potential. For example, current engines will retrieve and provide users with high aesthetic quality photos and guide aesthetic-driven image enhancement through aesthetic quality discriminators.Hence, it is desirable to automatically assess the aesthetic quality of the images.

Most previous works tried to use photography knowledge to guide the construction of artificial aesthetic features, such as the vivid colour, the rule of thirds, and the symmetry. However, these hand-craft and predefined features have limited representation ability, which is still a challenging task, although these features have shown encouraging results.

With the deep learning methods have shown great success in various computer vision tasks, more work has recently focused on using a deep convolutional neural network to extract effective aesthetics features. Image aesthetics assessment is typically cast as a classification or regression problem. Generally, the category or score

Further author information: (Send correspondence to Xiaodong Li)
Xiaodong Li : E-mail: lxdbesti@163.com

of the image will be used as indicators to distinguish the level of aesthetic quality, and these approaches often combined some other information, such as object, scene and attribute, to guide the task of aesthetic assessment. However, people all have different ideas about what is beautiful. In other words, the process of giving an aesthetic evaluation of a picture may be quite different for different people. Two pictures of the same score label may have different evaluation processes.

As Figure 1, shows six images and the associated aesthetic scores distribution from the AVA database[1] . Each line has the same score, but the aesthetic distribution is different. Although the higher the aesthetic score, the more attractive. However, the use of a single label cannot effectively express the potential difference of human aesthetics. Therefore, using a single label is not enough to reflect the aesthetic psychology of the user.In contrast, aesthetics distribution prediction is a more reasonable way to evaluate the diversified aesthetics of images. Some works use aesthetic rating distributions as ground-truth, and then use various loss functions such as Kullback-Leibler(KL) divergence, cumulative Jensen-Shannon divergence[2] or earth movers distance.[3]However, a most method is not considered the reason about the aesthetic distribution, more focus on reducing the distribution loss. In this paper, we propose a Deep Drift-Diffusion (DDD) model inspired by psychologists to predict aesthetic score distribution from images. The DDD model simulates various positive and negative attractors and a disturbance factor based on the deep image features so that the psychological processes among user can be effectively expressed. The experimental results in public aesthetic image datasets (AVA[1] and Photo.net[4]) reveal that our novel DDD model outperforms the state-of-the-art methods on aesthetic score distribution prediction.

In summary, our main contributions are as follows:

1. The first work that embeds drift-diffusion psychological model into deep convolutional neural networks;
2. The combination of a dynamic model in psychology and score distribution prediction of visual aesthetics;
3. Our work has the potential of inspiring more attentions to model the psychology process of aesthetic perception beyond just modeling of the aesthetic assessment results.

The main structure of this article is in the related work, we enumerate current work on aesthetic level assessment and aesthetic distribution evaluation. In the third chapter, we analyzed the statistical characteristics of the AVA dataset, and divided the data into various psychology model according to different math characteristics, and finally modeled different psychology models through deep learning. In Chapter 4, we compare the results of model with current methods for different aesthetic tasks. In Chapter 5, we summarized our work and put forward prospects for future work.

2. RELATED WORKS

2.1 Image Aesthetics Assessment

In the past decades, researchers on the topic of image aesthetic assessment have drawn much attention on 1D output: binary classification[4–7] and aesthetic scoring.[8] The binary classification is to give a binary label of image aesthetics: high or good and low or bad. The aesthetic scoring is to give a continues numerical score of image aesthetics: 0-1 or 0-10, the higher, the better.

However, only 1D binary label or aesthetic score can not fully describe the subjectiveness of aesthetic assessment. As claimed in Jin et al.,[2]an image with similar scores may differ in the score distributions, which are histograms of scores given by multiple image reviewers. The scores are the mean of the score distributions. The other statistics of the distribution such as variations, skewness, kurtosis can be quite different from the images with similar scores.

In the past decades, image aesthetics assessment was mainly through human aesthetic perception of image features and photography rules. The features included spatial distribution of edges, color distribution, hue and blur etc.[5] While drawing on some specific rules in photography, such as low depth-of-field indicator, then colorfulness measure, the shape convexity score and the familiarity measure,[4] The Rule of Thirds etc,[1] with the development of feature extraction technology, features based on high-level aesthetic principles had emerged, such as features based on scene types and related contents,[9] high-level semantic features based on this subject and background division.[10]

With the continuous advancement of deep learning research in recent years, CNN-based deep learning models were widely used in the classification and regression of aesthetics. Kao et al. proposed the multi-task learning.[6] They led the relevant items between tasks to the framework and made the utility of the appreciation of aesthetic and semantic labels more effective. Kong et al. and Chandakkar et al. utilized the relative aesthetics[8] to select the new datasets of image from the datasets with the pairs of related labels, and trained the related image pairs to obtain the aesthetic ranking with the higher accuracy. Sheng et al. adopted attention-based multi-patch aggregation to adjust the weight of each patch in the training process. The results improved learning effectively.[7]

2.2 Score distribution prediction

Recently, some methods are proposed to use modified or generated score distributions for binary classification and numerical assessment on aesthetics.[11–13] The pioneering work of Wu et al.[14] proposes a modified support vector regression algorithm to predict the score distribution in two small aesthetic datasets, before the large scale AVA dataset[1] released and the popularity of deep CNNs.

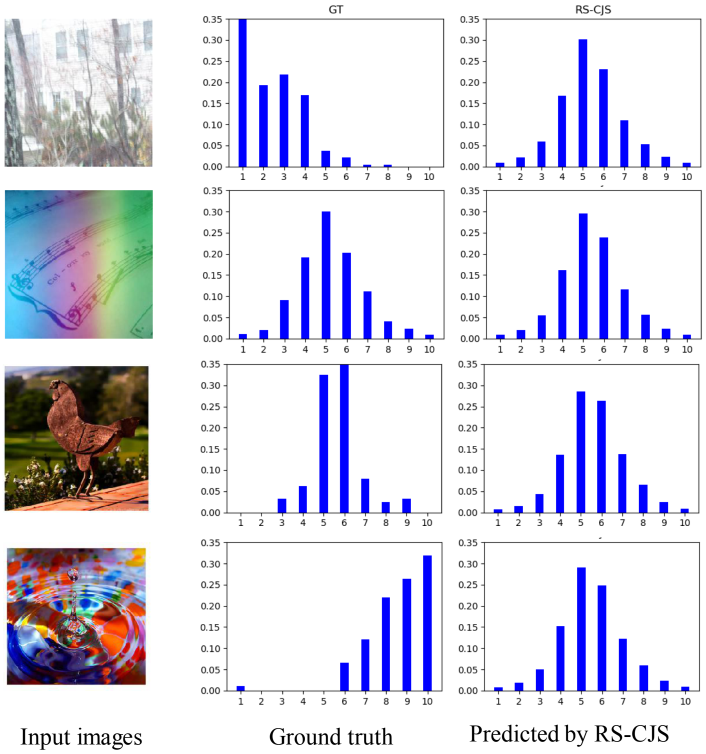

Figure 2. Examples of non Gaussian distributions of AVA ratings with low, middle and high scores. The state-of-the-art method RS-CJS can predict Gaussian like distributions well but fails in the two ends: the low and the high non Gaussian distributions. Even some distributions in the middle are non Gaussian such as the one shown in the 3rd line.

Most recently, Jin et al. propose a CNN based on the cumulative distribution with Jensen-Shannon divergence (RS-CJS)[2] to predict the aesthetic score distribution of human ratings, with a reliability-sensitive learning method based on the kurtosis of the score distribution. Talebi et al. propose a CNN based on EMD (Earth Mover's Distance) loss[3] to predict the aesthetic score distribution of human ratings. They use the predicted distribution to infer the mean score to guide the enhancement of images. Inspired by the human visual system, Xiaodan Zhang et al. propose the GPF-CNN architecture and GIF module.[15] The former can learn to focus on the important regions of the top-down neural attention map to extract fine detail features. The latter can adaptively fuse global and local features based on the input feature map. Gengyun Jia et al. consider that the aesthetic

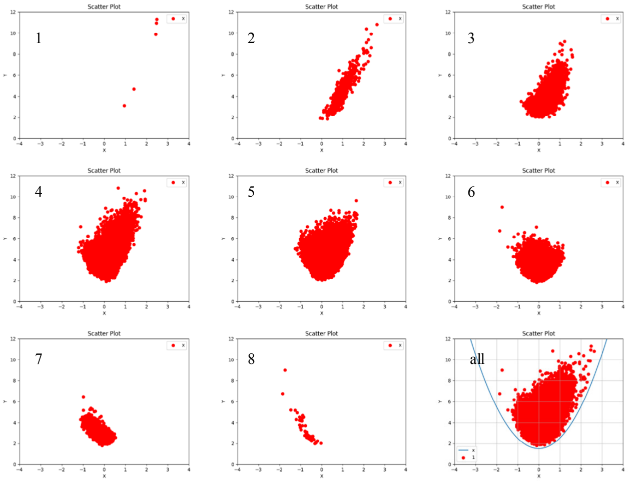

Figure 3. The skewness-kurtosis maps of AVA score distributions. In each sub-plot, the horizontal axis and the vertical axis are the skewness and the kurtosis, respectively. Each point represent an images in AVA. The number in each sub-plot indicate the score section of images. In the last sub-plot, we plot all the points together.

characteristics of images are a global feature,[16] and resizing pictures during model training will affect the aesthetic characteristics. So they were applying ROI pooling on feature maps of their aesthetic distribution model. Qiuyu Chen et al.[17] proposed an adaptive dilated convolution network to explicitly relate the aesthetic perception to the image aspect ratios while preserving the composition. Hui Zeng et al. create a comprehensive loss function to handle different aesthetic assessment tasks. Chaoran Cui et al.[18] develop a novel deep neural model that can enhance image aesthetic evaluation by collect information from object classification and scene recognition.

2.3 Subjectiveness of Aesthetics

The human ratings are quite subjective.[19] Unlike image recognition, people may give different scores of one image in the aspect of aesthetics. Chaoran Cui et al.[18] also draw this conclusion from the entropy of the AVA.[1] The Gaussian distribution is the best-performing model for only 62% of images in AVA. Examples of non-Gaussian distribution of human ratings are shown in Figure 2. The others are the skewed ones and can be best fitted by the Gamma distribution.[1]

The mean score is greatly influenced by the low and high extremes of the rating scale, which makes it inappropriate to be a robust estimation of the whole distribution, especially when the distribution is skewed. For skewed distributions, the median value appears to be more appropriate to describe the distributions than the mean value.[14]

Before our work, many methods have achieved good performance on score distributions, such as method of.[2,3,15,16,18] However, most of them follow the Gaussian's model. The mean score of the predicted distributions is always falling into [4:6]. This is because they have adopted a direct regression method. 62% of the distributions approximately follow Gaussian, which leads the regression results to be as similar as Gaussian distributions[2].[12] Thus we need not only learn the results of human ratings, but also find the underline processes of human ratings. The math models such as Gaussian or Gamma can not well model the aesthetic score distribution. We should find a psychological model that describe the processes of aesthetic perception.

Tae-Suh Park et al.[20] make a thoroughly analysis of the consensus of visual aesthetic perception. They use the skewness-kurtosis maps (S-K map) to measure and visualise the consensus of aesthetic scores.The function is defined as $K = S^2 + 1$. As shown in Figure 3, the skewness and kurtosis of a score distribution are the high moments compared with mean and variance. With the assistance of S-K maps, they find four patterns of aesthetic score distributions from AVA dataset. None of Gaussian or Gamma distribution can model the four patterns, especially the wide range of kurtosis. They propose to use a dynamic psychological processes model from psychology to model the processes of aesthetic perception not only the aesthetic assessment results. In the discussion part of the work of Tae-Suh Park et al.,[20] they hope that the future work will combine a dynamic model in psychology and score distribution prediction of visual aesthetics.

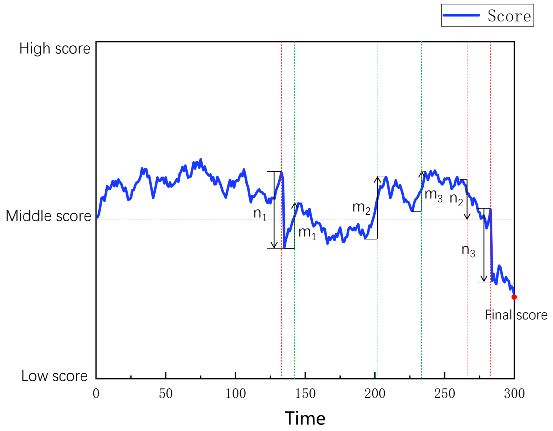

Figure 4. The simulation of the psychological processes of aesthetic perception of one rater. From the middle score (5 in AVA), the score will be disturbed by white noise. When a positive factor attracts the rater, the score will increase according to a exponential distribution, and vice verse. In this example, there are 3 positive attractors and 3 negative attractors. The final score is about 2.5.

3. THE PROPOSED DDD MODEL

3.1 Subjectiveness Analysis

Jin et al.[2] make a statistical analysis of subjectiveness or diversity of the opinion among annotators in a large-scale database for aesthetic visual analysis (AVA).[1] This dataset is specifically constructed for the purpose of learning more about image aesthetics. All those images are directly downloaded from dpchallenge.com. For each image in AVA, there is an associated distribution of scores (1-10) voted by different viewers. The number of votes that per image gets is ranged in 78-549, with an average of 210.

It is observed that most images' mean values are located in [4:7]. Images in this interval are not easy to be classified to a high-low label. Most images' standard deviation values are larger than 1:25, which shows the diversity of the human ratings for the same image. Images with mean score values from four to seven tend to have a low absolute value of the skewness and can be considered as those with symmetrical score distributions. Images with mean score values lower than four and greater than seven can be considered as those with positively and negatively skewed score distributions, respectively. This is likely due to the non-Gaussian nature of score distributions at the extremes of the rating scale. Within each range of the mean scores, there exist some images with high absolute values of kurtosis values (after normalized by minus 3), which are considered as those with unreliable score distributions.[2] Their exist wide range of kurtosis. This is the core reason that Gaussian distribution can not well fit the human ratings.

3.2 The Deep Drift-Diffusion Model

Inspired by Tae-Suh Park et al.,[20] we propose a Deep Drift-Diffusion (DDD) model based on psychology to predict the aesthetic score distribution of images. The DDD model combines the deep convolutional neural networks of artificial intelligence and the dynamic drift model from psychology[21–24] The DDD model simulates various positive and negative attractors and a disturbance factor based on the deep image features.

As shown in Figure 4, A person's aesthetic view is regarded as a process of being attracted by positive factors and influenced by negative factors. Suppose the initial score of a figure is 5 (5 in AVA),then the score will be disturbed by white noise. When a positive factor attracts the evaluators, the score will increase according to an exponential distribution. When the evaluator is affected by negative factors, the score will decrease according to an exponential distribution. The drift-diffusion model is described in Eq. 1.

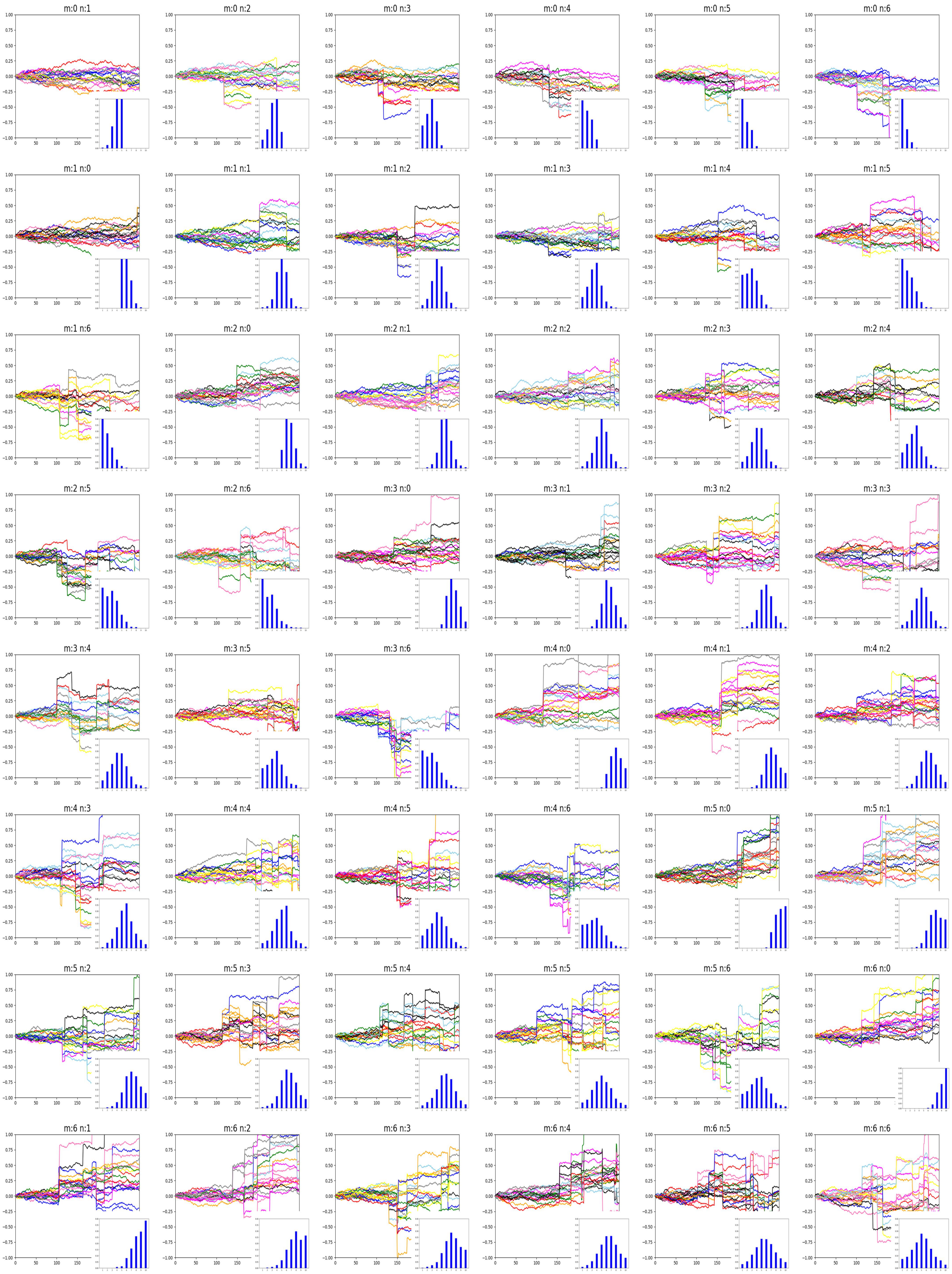

Figure 5. Score distributions simulated by n, m. $n, m \in [0, 6]$

The AVA dataset is from an online website. The photo reviews are given sufficient time to find the advantages and disadvantages of the aesthetics. Thus, the time factor does not that affect the results. Although there is no temporal component to the decision process in AVA dataset, the time factor in the process of psychological evaluation does not directly affect the outcome of the evaluation.

Where E_{pos} and E_{neg} follow exponential distribution (Eq. 2). W is our modified white noise as shown in Eq. 3. U is a uniform random distribution. Positive attractors represent the fluctuations of the score caused by good aesthetic factors, which can be simply interpreted as the advantages of the picture discovered by the viewer. While negative attractors are the opposite. The computational methods of E_{pos} and E_{neg} are consistent. Since they represent the same psychological status in the model, they must be characterised by the same distribution. According to this formula, one person's score is represented at a time of running. The score distribution can be recalculated by calculating the result of the formula for the number of times in the AVA images, thus generating training and testing labels. One example of 20 raters on one image with various ratios of m:n are shown in Figure 5.

$$v = MiddleScore + \sum_{i=1}^{m} E_{pos} - \sum_{i=1}^{n} E_{neg} + W, \tag{1}$$

$$E = 0.5 * e^{-0.5*U(0,10)} \tag{2}$$

$$W = 0.015 * U(-1, 1), \tag{3}$$

where the parameter 0.015 is from.[20]

We fit the ground truth score distributions using the Gaussian model and our DDD model respectively. The number of positive and negative attractors are determined by an exhaustive search. The ones who get the smallest distance between the simulated and the ground truth distributions are selected. The upper bound of the number of the positive or negative abstractors is 7, which is determined by experiments. The value above 7 does not decrease the fitting errors any more. Besides, In order to accurately calculate the exact positive and negative attractors, we enumerated the combination of all positive and negative attractors. Thus, we simulate 49 categories of psychological processes.

Through the combination of the number of positive and negative attractors, we can form 49 different specific quantitative relationships, which can represent 49 different psychological processes. And Tae-Suh Park and otherts[20] have only 4 psychological processes. Four examples of the fitted results are shown in Figure .6. Besides, we make numerical evaluation of the fitted results.

As shown in Table 1, in the middle, the Gaussian model wins our DDD a little. While at the two extreme ends, the DDD model can fit better than Gaussian model does. In addition, the DDD fits the four moments of the aesthetic distributions much better than the Gaussian does, as shown in Table 2.

The architecture of our DDD model is shown in Figure 7. We use a ResNet-50 as our basic model. Then we attach a multi-task regression module which contains the number of the positive (m) and the negative (n) attractors. The fitted model parameters are used as the ground truth labels for training. However, when evaluating our models in the experiments, we compare the distributions generated by our predicted parameters m and n with the original ground truth score distributions.

4. EXPERIMENTS

4.1 Datasets

To the best of our knowledge, the AVA dataset and the Photo.net *[4] dataset is the only two publicly available datasets with ground truth aesthetic score distributions. These two datasets are popular in the computational aesthetic community. We evaluate our DDD model on both AVA dataset[1] and Photo.net dataset.[4] We compare our models with the state-of-the-art methods of score distribution prediction task. Our DDD model outperforms all the state-of-the-art methods.

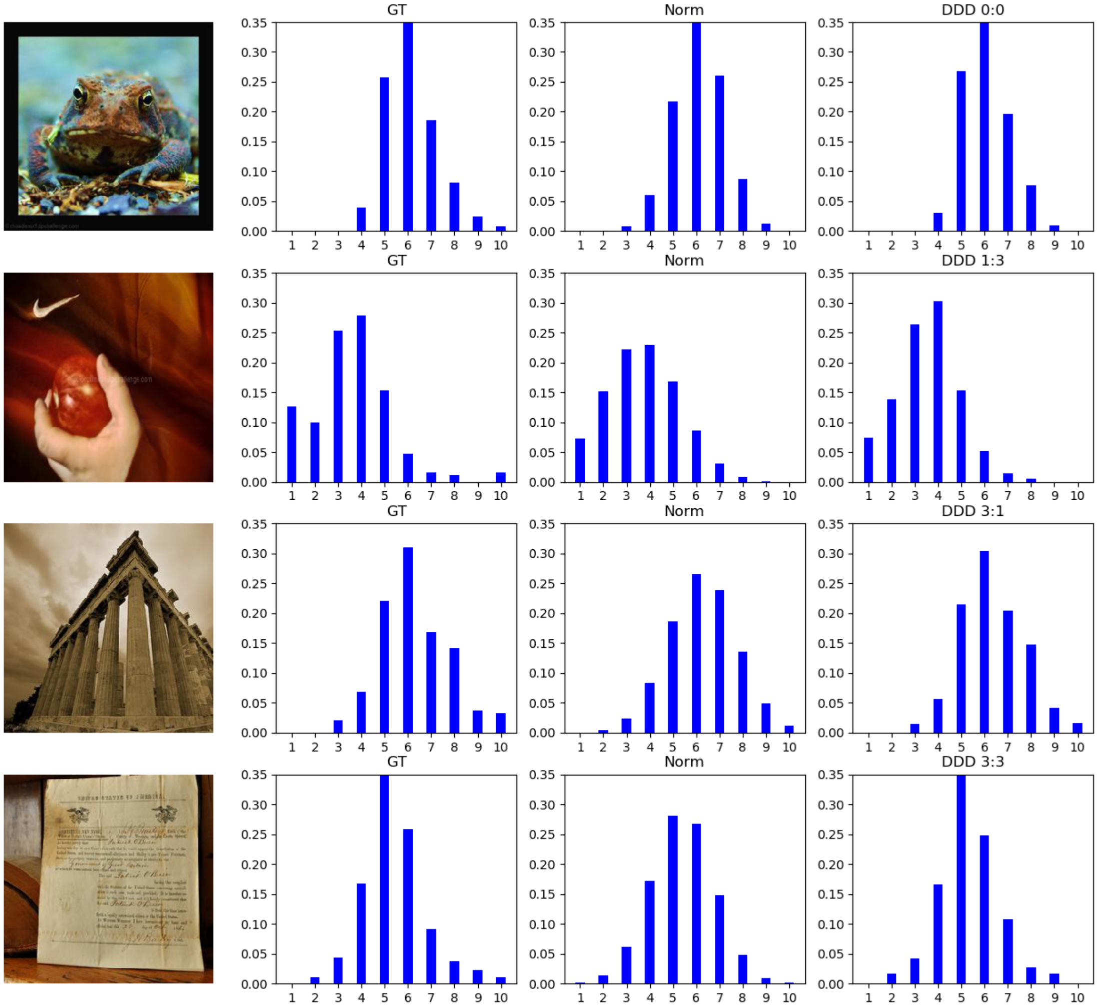

Figure 6. Examples of the fitting using Gaussian and our DDD model. Obviously, in the two extreme ends, the results of our DDD model are more similar as ground truth values. Images are from AVA dataset.

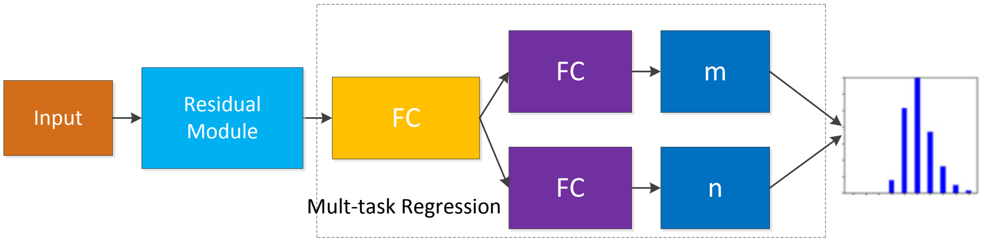

Figure 7. The architecture of the proposed deep drift-diffusion model. The ResNet-50 is our baseline CNN.

Table 1. The fitting errors by Gaussian and DDD, we use RMSE (Rooted Mean Square Error) to evaluate the errors.

Methods	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	all
Gaussian	0.117	0.063	0.030	**0.309**	**0.305**	**0.274**	0.276	0.524	0.0302
DDD Model	**0.026**	**0.030**	0.030	0.311	0.309	0.275	**0.265**	**0.310**	**0.0294**

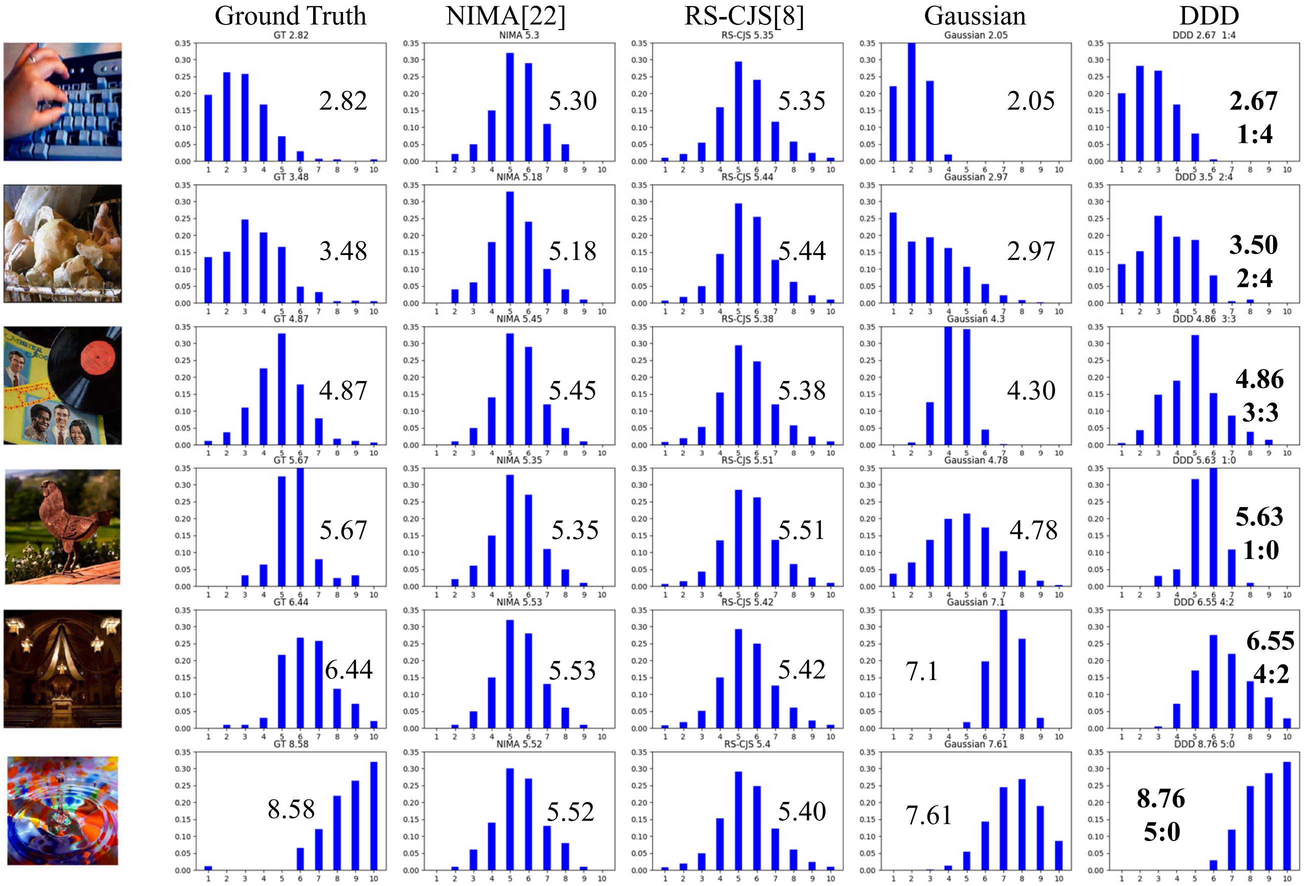

Figure 8. The comparison of aesthetic distribution prediction. The network used in this figure is ResNet-50. Images are from AVA dataset. Note that, different psychological processes can also be predicted by our DDD model with different ratios of m and n.

Table 2. The fitting errors of four moments by Gaussian and DDD, we use MSE (Mean Square Error) to evaluate the errors.

Methods	μ(MSE)	σ(MSE)	skew(MSE)	kurt(MSE)
Gaussian	≈0	0.0001	0.1951	1.0391
DDD Model	≈0	0.0001	**0.0805**	**0.4725**

Table 3. The comparison of aesthetic distribution prediction on AVA dataset. The lower the better.

Methods	PED	PCE	PJS	CED	CJS	PCS	PKL	EMD	Class. Acc. ↑
Gaussian(ResNet-101)	0.162	2.817	0.048	0.254	0.050	0.076	0.381	-	-
RS-CJS(1/3GoogleNet)[2]	0.158	2.760	0.037	0.260	0.040	0.068	0.323	-	80.08%
NIMA(inception-v2)[3]	0.168	2.693	0.028	0.137	0.029	0.044	0.081	0.050	81.51%
Hui Zeng et al.[25]	-	-	-	-	-	-	0.101	0.065	80.81%
Chaoran Cui et al.[18]	0.127	-	-	-	-	-	0.094	-	-
Gengyun Jia et al.[16]	-	-	-	-	-	-	-	0.041	-
Xiaodan Zhang et al.[15]	-	-	-	-	-	-	-	0.045	81.81%
DDD(1/3GoogLeNet)	0.142	2.729	0.028	0.177	0.026	0.051	0.153	0.044	81.59%
DDD(MobileNet-v1)	0.125	2.669	0.022	0.138	0.019	0.042	0.092	0.031	80.43%
DDD(ResNet-50)	0.109	2.667	0.020	0.129	0.015	0.035	0.071	0.026	82.63%
DDD(ResNet-101)	**0.105**	**2.640**	**0.019**	**0.122**	**0.013**	**0.028**	**0.065**	**0.023**	**82.65%**

Table 4. The comparison of aesthetic distribution prediction on Photo.net dataset. The lower the better.

Methods	PED	PCE	PJS	CED	CJS	PCS	PKL	EMD	Class. Acc. ↑
Gaussian(1/3GoogLeNet)	0.313	2.351	0.097	0.348	0.066	0.179	1.432	0.075	73.54%
RS-CJS(1/3GoogLeNet)[2]	0.305	2.270	0.085	0.311	0.060	0.143	1.340	0.072	75.62%
DDD(1/3GoogLeNet)	0.289	2.208	0.073	0.260	0.054	0.121	1.247	0.070	77.96%
Gaussian(ResNet-50)	0.296	2.164	0.093	0.292	0.064	0.153	1.273	0.073	76.76%
RS-CJS(ResNet-50)[2]	0.262	1.963	0.071	0.264	0.059	0.138	1.185	0.069	78.10%
DDD(ResNet-50)	0.243	1.842	0.068	0.251	0.106	0.035	1.129	0.066	79.22%
Gaussian(ResNet-101)	0.293	2.157	0.092	0.289	0.061	0.149	1.268	0.071	76.84%
RS-CJS(ResNet-101)[2]	0.255	1.961	0.070	0.262	0.058	0.137	1.183	0.068	78.12%
DDD(ResNet-101)	**0.242**	**1.840**	**0.068**	**0.249**	**0.047**	**0.105**	**1.126**	**0.064**	**79.26%**

AVA. The AVA dataset is a list of image ids from DPChallenge.com, which is an online photography social network. There are total 255,530 photographs, each of which is rated by 78–549 persons, with an average of 210 aesthetic ratings ranging from 1 to 10. We follow the standard partition method of the AVA dataset in previous work.[1,2,8,26–28] The training and testing sets contain 235,599 and 19,930 images respectively.

Photo.net. Each image in the Photo.net dataset is rated by at least ten users to evaluate the aesthetic quality from 1 to 7. Due to some unavailable links in photo.net website, we collect 15,582 images in all. We follow the partition ratio in previous work.[4,29] The training and testing sets contain 13,582 and 2000 images, respectively. For the aesthetic quality classification task, we also follow[4,29] and choose the average score of 5.0 as median aesthetic ratings. The images with an average score larger than $5 + \delta$ are designated as high quality images, those with an average score smaller than 5 as low-quality images. We set δ to 0 in the experiment, which is more challenging than that with setting δ to other values.[1]

4.2 Implementation Details

We fix the parameters of the layers before the first fully connected layer of a pre-trained GoogLeNet model and ResNet mode1[30] on the ImageNet and fine-tune the all full connected layers on the training set of the AVA dataset. We use the Caffe framework to train and test our models. The learning policy is set to step. Stochastic gradient descent is used to train our model with a mini-batch size of 48 images, a momentum of 0.9, a gamma of 0.5 and a weight decay of 0.0005. The max number of iterations is 120000. The training time is about 5 hours and 8 hours using Titan X Pascal GPU.

[*] http://ritendra.weebly.com/aesthetics-datasets.html

4.3 Score Distribution Prediction

We compare our DDD model with the method of RS-CJS,[2] which uses 1/3 GoogLeNet and the RS-CJS loss. The evaluation rules follow those in.[2] The numerical results are shown in Table 3 (AVA) and Table 4 (Photo.net). For a fair comparison, we also use 1/3 GoogLeNet to replace the ResNet-50 in Figure 7. Besides, we modify the regression targets of our DDD to the μ and σ of the fitted Gaussian distributions. The numerical results in Table 3 and Table 4 reveal that our DDD model beats the Gaussian model and the RS-CJS, which directly fit the results no matter using 1/3 GoogLeNet, ResNet-50 or ResNet-101. The performances of the Gaussian model are even worse than RS-CJS.[2]

We also give some visualized comparison results in Figure 8. Scores of most images of AVA dataset are in the range of [4,6]. Thus the NIMA[3] and RS-CJS[2] tend to output distributions in the middle. The mean scores of their distributions also fall in [4,6]. The regression errors are less influenced by the image with very low or very high scores. The Gaussian distribution can not fit the original distribution well. The predicted distributions of our DDD model fit well the ground truth distributions. We can not only get output middle scores but also low and high scores. Besides, the images with middle scores can be divided by the different ratios of m and n of the DDD model, which represent different kinds of psychological processes.

As shown in Table 3 and Table 4, our DDD model outperforms all the state-of-the-art methods[2, 3, 15, 16, 18, 25] on the aesthetic score distribution prediction task using the evaluation metrics of,[2] which contain 8 different kinds of distribution distances. All the previous state-of-the-art methods do not take the psychological process of aesthetics into consideration. They only model the aesthetic perception results in the form of score distributions of multiple reviewers.

4.4 Aesthetic Classification

We recast our predicted score distribution to 1-dimensional binary label and compare with the state-of-the-art methods on aesthetic quality classification, as shown in the last column of Table 3. Compared to the state-of-the-art method on aesthetic score distribution prediction,[2, 3, 15, 25] our DDD model achieves the state-of-the-art aesthetic classification on AVA dataset. Note that, on the AVA dataset, our DDD model with only 1/3 GoogleNet beats NIMA model[3] with full Inception-v2 GoogleNet on the aesthetic classification task.

5. CONCLUSIONS AND DISCUSSIONS

In this paper, we propose a DDD model inspired by psychologists to predict aesthetic score distribution from images. The DDD model simulates various positive and negative attractors and a disturbance factor based on the deep image features. The experimental results in large scale aesthetic image datasets (AVA and Photo.net) reveal that our novel DDD model outperforms the state-of-the-art methods in aesthetic score distribution prediction. Besides, different psychological processes can also be predicted by our model.

The original drift diffusion psychology model has a temporal component of the decision process. In future work, we will build such an aesthetic dataset with score distributions and rating time data, which models the dynamic process of the aesthetic perception. Besides, we will explore more psychological processes of aesthetic perception.

REFERENCES

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

[2] Jin, X., Wu, L., Li, X., Chen, S., Peng, S., Chi, J., Ge, S., Song, C., and Zhao, G., "Predicting aesthetic score distribution through cumulative jensen-shannon divergence," in [*Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*], (2018).

[3] Talebi, H. and Milanfar, P., "NIMA: neural image assessment," *IEEE Trans. Image Processing* **27**(8), 3998–4011 (2018).

[4] Datta, R., Joshi, D., Li, J., and Wang, J. Z., "Studying aesthetics in photographic images using a computational approach," in [*ECCV, Graz, Austria, May 7-13, 2006, Proceedings, Part III*], 288–301 (2006).

[5] Luo, W., Wang, X., and Tang, X., "Content-based photo quality assessment," in [*2011 International Conference on Computer Vision*], 2206–2213, IEEE (2011).

[6] Kao, Y., He, R., and Huang, K., "Deep aesthetic quality assessment with semantic information," *IEEE Transactions on Image Processing* **26**(3), 1482–1495 (2017).

[7] Sheng, K., Dong, W., Ma, C., Mei, X., Huang, F., and Hu, B.-G., "Attention-based multi-patch aggregation for image aesthetic assessment," in [*2018 ACM Multimedia Conference on Multimedia Conference*], 879–886, ACM (2018).

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

[9] Dhar, S., Ordonez, V., and Berg, T. L., "High level describable attributes for predicting aesthetics and interestingness," in [*CVPR 2011*], 1657–1664, IEEE (2011).

[10] Luo, Y. and Tang, X., "Photo and video quality evaluation: Focusing on the subject," in [*European Conference on Computer Vision*], 386–399, Springer (2008).

[11] Jin, B., Segovia, M. V. O., and Süsstrunk, S., "Image aesthetic predictors based on weighted cnns," in [*2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*], 2291–2295 (2016).

[12] Wang, Z., Liu, D., Chang, S., Dolcos, F., Beck, D., and Huang, T. S., "Image aesthetics assessment using deep chatterjee's machine," in [*IJCNN*], 941–948, IEEE (2017).

[13] Hou, L., Yu, C., and Samaras, D., "Squared earth mover's distance-based loss for training deep neural networks," *CoRR* **abs/1611.05916** (2016).

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

[15] Zhang, X., Gao, X., Lu, W., and He, L., "A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction," *IEEE Transactions on Multimedia* (2019).

[16] Jia, G., Li, P., and He, R., "Theme aware aesthetic distribution prediction with full resolution photos," *arXiv preprint arXiv:1908.01308* (2019).

[17] Chen, Q., Zhang, W., Zhou, N., Lei, P., Xu, Y., Zheng, Y., and Fan, J., "Adaptive fractional dilated convolution network for image aesthetics assessment," *arXiv preprint arXiv:2004.03015* (2020).

[18] Cui, C., Liu, H., Lian, T., Nie, L., Zhu, L., and Yin, Y., "Distribution-oriented aesthetics assessment with semantic-aware hybrid network," *IEEE Transactions on Multimedia* **21**(5), 1209–1220 (2018).

[19] Kim, W. H., Choi, J. H., and Lee, J. S., "Objectivity and subjectivity in aesthetic quality assessment of digital photographs," *IEEE Transactions on Affective Computing* , 1–1 (2018).

[20] Park, T. and Zhang, B., "Consensus analysis and modeling of visual aesthetic perception," *IEEE Trans. Affective Computing* **6**(3), 272–285 (2015).

[21] Kelso, J. S. et al., "The self-organization of brain and behavior," *Lecture Notes in Complex Systems, Santa Fe, NM* (1995).

[22] Ratcliff, R., "A theory of memory retrieval.," *Psychological review* **85**(2), 59 (1978).

[23] Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D., "The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks.," *Psychological review* **113**(4), 700 (2006).

[24] Ratcliff, R. and McKoon, G., "The diffusion decision model: theory and data for two-choice decision tasks," *Neural computation* **20**(4), 873–922 (2008).

[25] Zeng, H., Cao, Z., Zhang, L., and Bovik, A. C., "A unified probabilistic formulation of image aesthetic assessment," *IEEE Transactions on Image Processing* **29**, 1548–1561 (2019).

[26] Wang, W., Zhao, M., Wang, L., Huang, J., Cai, C., and Xu, X., "A multi-scene deep learning model for image aesthetic evaluation," *Sig. Proc.: Image Comm.* **47**, 511–518 (2016).

[27] Lu, X., Lin, Z. L., Jin, H., Yang, J., and Wang, J. Z., "Rating image aesthetics using deep learning," *IEEE Trans. Multimedia* **17**(11), 2021–2034 (2015).

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

[29] Kao, Y., He, R., and Huang, K., "Deep aesthetic quality assessment with semantic information," *IEEE Trans. Image Processing* **26**(3), 1482–1495 (2017).

[30] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*CVPR*], 770–778, IEEE Computer Society (2016).