Image Aesthetics Assessment using Deep Chatterjee's Machine

Zhangyang Wang†, Ding Liu‡, Shiyu Chang◇, Florin Dolcos‡, Diane Beck‡, Thomas Huang‡
†Department of Computer Science and Engineering, Texas A&M University, College Station, TX
‡Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL
◇ IBM Thomas J. Watson Research Center, Yorktown Heights, NY
Email: *atlaswang@tamu.edu, {dingliu2, fdolcos, dmbeck, t-huang1}@illinois.edu, shiyu.chang@ibm.com*

***Abstract*—Image aesthetics assessment has been challenging due to its subjective nature. Inspired by the Chatterjee's visual neuroscience model, we design Deep Chatterjee's Machine (DCM) tailored for this task. DCM first learns attributes through the parallel supervised pathways, on a variety of selected feature dimensions. A high-level synthesis network is trained to associate and transform those attributes into the overall aesthetics rating. We then extend DCM to predicting the distribution of human ratings, since aesthetics ratings are often subjective. We also highlight our first-of-its-kind study of label-preserving transformations in the context of aesthetics assessment, which leads to an effective data augmentation approach. Experimental results on the AVA dataset show that DCM gains significant performance improvement, compared to other state-of-the-art models.**

I. Introduction

Automated assessment or rating of pictorial aesthetics has many applications, such as in an image retrieval system or a picture editing software [1]. Compared to many other typical machine vision problems, the aesthetics assessment is even more challenging, due to the highly subjective nature of aesthetics, and the seemingly inherent semantic gap between low-level computable features and high-level human-oriented semantics. Though aesthetics influences many human judgments, our understanding of what makes an image aesthetically pleasing is still limited. Contrary to semantics, an aesthetics response is usually very subjective and difficult to gauge even among human beings.

Existing research has predominantly focused on constructing hand-crafted features that are empirically related to aesthetics. Those features are designed under the guidance of photography and psychological rules, such as rule-of-thirds composition, depth of field (DOF), and colorfulness [2], [3]. With the images being represented by these hand-crafted features, aesthetic classification or regression models can be trained on datasets consisting of images associated with human aesthetic ratings. However, the effectiveness of hand-crafted features is only empirical, due to the vagueness of certain photographic or psychologic rules. Recently, deep learning [4] has achieved prevailing success, ranging from object recognition [5], to the more subtle and subjective style recognition [6], the latter of which bears certain connections to the assessment of aesthetics. Lu et.al. [7] proposed the *Rating Pictorial Aesthetics using Deep Learning* (**RAPID**) model, with impressive accuracies on the *Aesthetic Visual Analysis* (**AVA**) dataset [8]. However, they have not yet studied more precise predictions, such as finer-grain ratings or rating distributions [9]. On the other hand, the study of the cognitive and neural underpinnings of aesthetic appreciation by means of neuroimaging techniques yields some promise for understanding human aesthetics [10]. Although the results of these studies have been somewhat divergent, a hierarchical set of core mechanisms involved in aesthetic preference have been identified [11].

In this work, we develop a novel deep-learning based image aesthetics assessment model, called *Deep Chatterjee's Machine* (**DCM**). DCM clearly distinguishes itself from prior models, for its unique architecture inspired the Chatterjee's visual neuroscience model [12]. We introduce the specific architecture of *parallel supervised pathways*, to learn multiple attributes on a variety of selected feature dimensions. Those attributes are then associated and transformed into the overall aesthetic rating, by a *high-level synthesis network*. Our technical contribution also includes the study of label-preserving transformations in the context of aesthetics assessment, which is applied to effective data augmentation. We examine DCM on the large-scale AVA dataset [8], for the aesthetics rating prediction task, and confirms its superiority over a few competitive methods, with the same or larger amounts of parameters.

A. Related Work

Datta et.al. [2] first casted the image aesthetics assessment problem as a classification or regression problem. A given image is mapped to an aesthetic rating, which is usually collected from multiple subject raters and is normally quantized with discrete values. [2], [3] extracted various handcrafted features, including low-level image statistics such as distributions of edges and color histograms, and high-level photographic rules such as the rule of thirds. A few subsequent efforts, such as [13], [14], [15], focus on improving the quality of those features. Generic image features [16], such as SIFT and Fisher Vector [17], were applied to predict aesthetics. However, empirical features cannot accurately and exhaustively represent the aesthetic properties.

978-1-5090-6182-2/17/$31.00 ©2017 IEEE

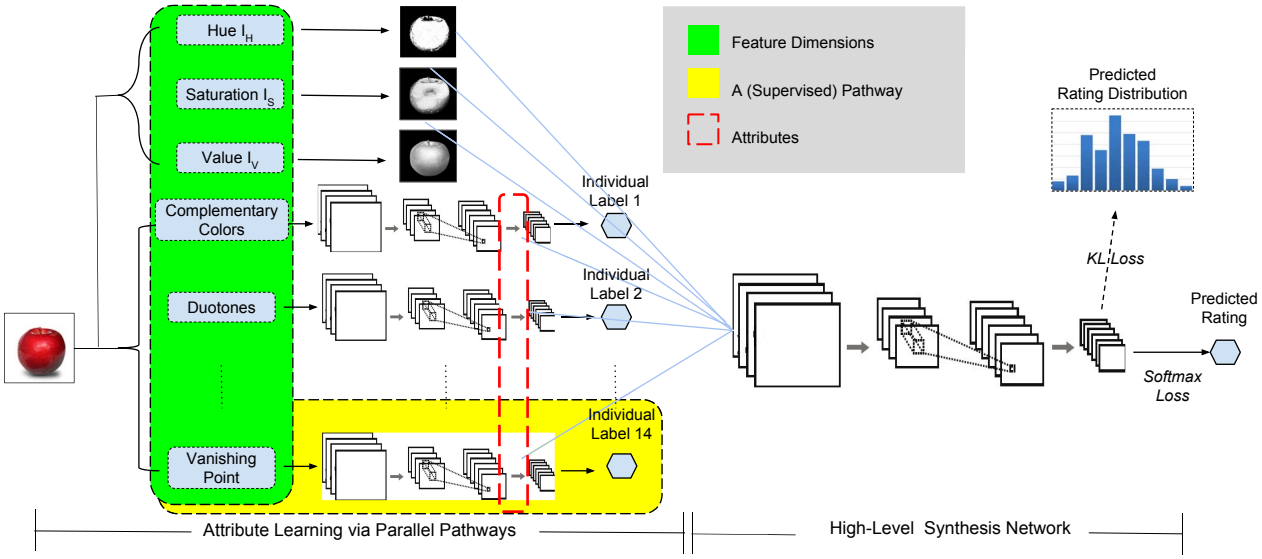


Fig. 1. The architecture of Deep Chatterjee's Machine (DCM). The input image is first processed by parallel pathways, each of which learns an attribute along a selected feature dimension independently. Except for the first three simplest features (*hue, saturation, value*), all parallel pathways take the form of fully-convolutional networks, supervised by individual labels; their hidden layer activations are utilized as learned attributes. We then associate those *"pre-trained"* pathways with the high-level synthesis network, and *jointly tune* the entire network to predict the overall aesthetics ratings.

The human brain transforms and synthesizes a torrent of complex and ambiguous sensory information into coherent thought and decisions. Most aesthetic assessment methods adopt simple linear classifiers to categorize the input features, which is obviously oversimplified. Deep networks [18] attempt to emulate the underlying complex neural mechanisms of human perception, and display the ability to describe image content from the primitive level (low-level features) to the abstract level (high-level features). The RAPID model [7] is among the first to apply deep convolutional neural networks (CNN) [4] to the aesthetics rating prediction, where the features are automatically learned. They further improved the model by exploring style annotations [8] associated with images. In fact, even the hidden activations from a generic CNN was found to work reasonably well for aesthetics features [19].

Most current work treat aesthetics assessment as a conventional classification problem: the user ratings of each photo are transformed into a ordinal scalar rating (by averaging, etc.), which is taken as the label of this photo. For example, RAPID [7] divided all samples as aesthetic or unaesthetic, and trained a binary classification model. Contrary to the oversimplified setting, it is common for different users to rate the same visual subject inconsistently or even oppositely, due to the subjective problem nature [3]. Since human aesthetic assessment depends on multiple dimensions such as composition, colorfulness, or even emotion [20], it is difficult for individuals to reliably convert their experiences to a single rating, resulting in noisy estimates of real aesthetic responses. In [9], Wu et.al. proposed to represent each photo's rating as a distribution vector over basic ratings, constituting a structural regression problem. A multi-label aesthetic assessment task was discussed in [21], where aesthetic attributes were predicted jointly

B. Datasets

Large and reliable datasets, consisting of images and corresponding human ratings, are the essential foundation for the development of machine assessment models. Several Web photo resources have taken advantage of crowdsourcing contributions, such as Flickr and DPChallenge.com [8]. The AVA dataset is a large-scale collection of images and meta-data derived from DPChallenge.com. It contains over 250,000 images with aesthetic ratings from 1 to 10, and a 14,079 subset with binary style labels (e.g., rule of thirds, motion blur, and complementary colors), making automatic feature learning using deep learning approaches possible. In this paper, we focus on AVA as our research subject.

II. The Neuroaesthetics Models

Multiple parallel processing strategies, involving over a dozen retinal ganglion cell types, can be found in the retina. Each ganglion cell type tiles the retina to focus on one specific kind of feature, and provide a complete representation across the entire visual field [22]. Retinal ganglion cells project in parallel from the retina, through the lateral geniculate nucleus of the thalamus to the primary visual cortex. Primary visual cortex receives parallel inputs from the thalamus and uses modularity, defined spatially and by cell-type specific connectivity, to recombine these inputs into new parallel outputs. Beyond primary visual cortex, separate but interacting dorsal and ventral streams perform distinct computations on similar visual information to support distinct behavioural goals [23]. The integration of visual information is then achieved progressively. Independent groups of cells with different functions are brought into temporary association, by a so-called "binding" mechanism [10], for the final decision-making.

From the retina to the prefrontal cortex, the human visual processing system will first conduct a very rapid holistic image analysis [24], [25], [26]. The divergence comes at a later stage, in how the low-level visual features are further processed through parallel pathways [27] before being utilized. The pathway can be characterized by a hierarchical architecture, in which neurons in higher areas code for progressively more complex representations by pooling information from lower areas. For example, there is evidence [28] that neurons in V1 code for relatively simple features such as local contours and colors, whereas neurons in TE fire in response to more abstractive features, that encode the scene's gist and/or saliency information and act as a holistic signature of the input.

Key Notations: For the consistency of terms, we use *feature dimension* to denote a prominent visual property, that is relevant to aesthetics judgement. We define an *attribute* as the learned abstracted, holistic feature representation over a specific feature dimension. We define a *pathway* as the processing mechanism from a raw visual input to an attribute.

A. Chatterjee's Visual Neuroscience Model

The main insights for DCM were gained from the classical and important **Chatterjee's visual neuroscience model** [12]. It models the cognitive and affective processes involved in visual aesthetic preference, providing a means to organize the results obtained in the 2004-2006 neuroimaging studies, within a series of information-processing phases. The Chatterjee's model concludes the following simplified, but important insights, that inspire our model:

- The human brain works as a multi-leveled system.
- For the visual sensory input, a variety of relevant feature dimensions are first targeted.
- A set of parallel pathways abstract the visual input. Each pathway processes the input into an attribute on a specific feature dimension.
- The high-level association and synthesis transforms all attributes into an aesthetics decision.

Step 2 and 3 are derived from the many recent advances [22] showing that aesthetics judgments evidently involve multiple pathways, which could connect from related perception tasks [10], [11]. Previously, many feature dimensions, such as color, shape, and composition, have already been discovered to be crucial for aesthetics. A bold yet rational assumption is thus made by us, that the attribute learning for aesthetics tasks could be decomposed onto those pre-known feature dimensions and processed in parallel.

III. Deep Chatterjee's Machine

The architecture of Deep Chatterjee's Machine (DCM) is depicted in Fig. 1. The whole training process is divided in two stages, based on the above insights. In brief, we first learn attributes through parallel (supervised) pathways, over the selected feature dimensions. We then combine those *"pre-trained"* pathways with the high-level synthesis network, and *jointly tune* the entire network to predict the overall aesthetics ratings. The testing process is completely feed-forward and end-to-end.

TABLE I
The 14 style attribute annotations in the AVA dataset

Style	Number	Style	Number
Complementary Colors	949	Duotones	1, 301
High Dynamic Range	396	Image Grain	840
Light on White	1,199	Long Exposure	845
Macro	1,698	Motion Blur	609
Negative Image	959	Rule of Thirds	1,031
Shallow DOF	710	Silhouettes	1,389
Soft Focus	1,479	Vanishing Point	674

A. Attribute Learning via Parallel Pathways

1) Selecting Feature Dimensions: We first select feature dimensions that are discovered to be highly related to aesthetics assessment. Despite the lack of firm rules, certain visual features are believed to please humans more than others [2]. We take advantage of those photographically or psychologically inspired features as priors, and force DCM to "focus" on them.

The previous work, e.g., [2]. has identified a set of aesthetically discriminative features. It suggested that the light exposure, saturation and hue play indispensable roles. We assume the RGB data of each image is converted to HSV color space, as I_H, I_S, and I_V, where each of them has the same size as the original image[1]. Furthermore, many photographic style features influence human's aesthetic judgements. [2] proposed six sets of photographic styles, including the rule of thirds composition, textures, shapes, and shallow depth-of-field (DOF). The AVA dataset comes with a more enriched variety of *style annotations*, as listed in Table I, which are leveraged by us. [2]

2) Parallel Supervised Pathways: Among the 17 feature dimensions, the simplest three, I_H, I_S, and I_V are immediately obtained from the input. However, the remaining 14 style feature dimensions are not qualitatively well-defined; their attributes are not straightforward to be extracted.

For each style category as a feature dimension, we create binary *individual labels*, by labelling images with the style annotation as "1" and otherwise "0", which follows many previous work [8], [14]. We design a special architecture, called *parallel supervised pathways*. Each pathway is modeled with a *fully convolutional neural network* (FCNN), as in Fig. 2. It takes an image as the input, and outputs image's individual label along this feature dimension. All pathways are learned in parallel without intervening with each other. The choice of FCNN is motivated by the spatial locality-preserving property of human brain's low-level visual perception [23].

[1]We downsample I_H, I_S, and I_V to 1/4 of their original size, to improve the efficiency. It turns out that the model performance is hardly affected, which is understandable since the human perceptions of those features are insensitive to scale changes.

[2]The 14 photographic styles are chosen specifically on the AVA datasets. We do not think they represent all aesthetics-related visual information, and plan to have more photographic styles annotated.

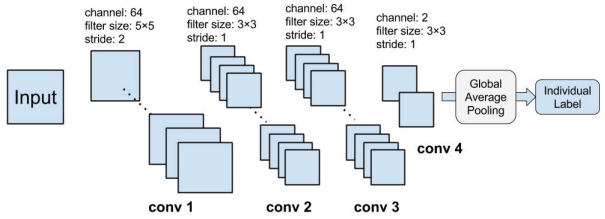


Fig. 2. The architecture of a supervised pathway as a FCNN. A 2-way softmax classifier is employed after global averaging pooling, to predict the individual label (0 or 1).

For each feature dimension, the number of labeled samples is limited, as shown in Table I. Therefore, we pre-train the first two layers in Fig. 2, using all images from the AVA dataset, in a unsupervised way. We construct a 4-layer Stacked Convolutional Auto Encoder (SCAE): its first 2 layers follows the same topology as the conv1 and conv2 layers, and the last 2 layers are mirror-symmetrical deconvolutional layers [29]. After SCAE is trained, the first two layers are applied to initialize the conv1 and conv2 layers for all 14 FCNN pathways. The strategy is based on the common belief that the lower layers of CNNs learn general-purpose features, such as edges and contours, which could be adapted for extensive high-level tasks [30]. After the initialization of the first two layers, for each pathway, we concatenate them to the conv3 and conv4 layers, and further conduct supervised training using individual labels. The conv4 layer always has the same channel number with the corresponding style classes (here 2 for all, since we only have binary labels for each class). It is followed by the global average pooling [31] step, to be correlated with the binary labels. Eventually, the conv4 layer as well as the classifier are discarded, and the conv1-conv3 layers of 14 pathways are passed to the next stage. We treat the conv3 layer activations of each pathway as learned attributes [30].

The pathways in DCM accounts for progressively extracting more complex features. As observed in experiments, the pre-training of all pathways' conv1 and conv2 layers learns shared low-level features, such as edges and blobs. Each pathway is then independently tuned by its "higher-level" concepts, which guides the adaption of low-level features. The final outputs of pathways, conv3, are abstracted from the low-level conv1 and conv2 features, and are regarded as mid-level attributes. Each pathway's conv3 attribute displays a different, visible combination of low-level features, but not any semantically meaningful object.

B. Training High-Level Synthesis Network

Finally, we simulates brain's high-level association and synthesis, using a larger FCNN. Its architecture resembles Fig. 2, except that the first three convolutional layers each have 128 channels instead of 64. The high-level synthesis network takes the attributes from all parallel pathways as inputs, and outputs the overall aesthetics rating. The entire DCM is then tuned from end to end.

IV. Predicting The Distribution Representation

Most existing studies [2] apply a scalar value to represent the predicted aesthetics quality, which appears insufficient to capture the true subjective nature. For example, two images with the equal mean score could have very different deviations among raters. Typically, an image with a large rating variance is more likely to be edgy or subject to interpretation. [7] assigned images with binary aesthetics labels, i.e., high quality and low quality, by thresholding their mean ratings, which provided less informative supervision due to the large intra-class variation. [9] suggested to represent the ratings as a distribution on pre-defined ordinal basic ratings. However, such a structural label could be very noisy, due to the coarse grid of basic ratings, the limited sample size (number of ratings) per image, and the lack of shifting robustness of their L_2-based loss.

The previous study of the AVA datasets [8] reveals two important facts:

- For all images, the standard deviation of an image's ratings is a function of its mean rating. Especially, images with "moderate" ratings tend to have a lower variance than images with "extreme" ratings. It inspires us that the estimations of mean ratings and standard deviations may be jointly performed, which can potentially mutually reinforce each other.
- For each image, the distribution of its ratings from different raters is largely Gaussian. According to [8], Gaussian functions perform adequately good approximations to fit the rating distributions of 99.77% AVA images. Besides, those non-Gaussian distributions tend to be highly-skewed, occurring at the low and high extremes of the rating scale, where their mean ratings could be predicted with higher confidences.

To this end, we propose to explicitly model the rating distribution for each image as Gaussian, and jointly predict its mean and standard deviation. Assuming the underlying distribution $N_1(\mu_1, \sigma_1)$ and the predicted distribution $N_2(\mu_2, \sigma_2)$, their difference is calculated by the Kullback-Leibler (KL) divergence [32]:

$$KL(N_1, N_2) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\mu_2^2} - \frac{1}{2} \qquad (1)$$

N_1 is calculated by fitting the rating histogram (over the 10 discrete ratings) of each image, with a Gaussian model. It is treated as the "ground truth" here. $KL(N_1, N_2) = 0$ if and only if the two distributions are exactly the same, and increases while N_2 diverges from N_1.

When training DCM to predict rating distributions, we replace the default softmax loss with the loss function (IV), which corresponds to the KL-loss branch (the dash) in Fig. 1. The outputs of the global average pooling from the high-level synthesis network remains to be a vector $\in R^{2\times 1}$. But different from the binary prediction task where the output denotes a Bernoulli distribution over [0, 1] labels, the two elements in the output here denote the predicted mean and variance, respectively. They could thus be arbitrary real values falling within the rating scale.

(a) (b)

Fig. 3. DCM classification examples: (a) high-quality; (b) low-quality (δ = 0).

V. Study Label-Preserving Transformations

When training deep networks, the most common approach to reduce overfitting is to artificially enlarge the dataset using label-preserving transformations [32]. In [4], image translations and horizontal reflections are generated, while the intensities of the RGB channels are altered, both of which apparently will not change the object class labels. Other alternatives, such as random noise, rotations, warping and scaling, are also widely adopted by the latest deep-learning based object recognition methods. However, there has been little work on identifying label-preserving transformations for image aesthetics assessment, e.g., those that will not significantly alter the human aesthetics judgements, considering the rating-based labels are very subjective. In [7], motivated by their need to create fixed-size inputs, the authors created randomly-cropped local regions from training images, which was empirically treated as data augmentation.

We make the first exploration to identify whether a certain transformation will preserve the *binary aesthetics rating*, i.e., high quality versus low quality, by conducting a **subjective evaluation survey** among over 50 participants. We select 20 high-quality (δ = 1) images from the AVA dataset (since low-quality images are unlikely to become more aesthetically pleasing after some simple/random transformations). Each image is processed by all different kinds of transformations in Table II. For each time, a participant is shown with a set of image pairs originated from the same image, but processed with different transformations. The groundtruth is also included in the comparison process. For each pair, the participant needs to decide which one is better in terms of aesthetics quality. The image pairs are drawn randomly, and the image winning this pairwise comparison will be compared again in the next round, until the best one is selected.

We fit a Bradley-Terry model [33] model to estimate the subjective scores for each method so that they can be ranked, which is similar to [34]. With the groundtruth set as score 1, each transformation will receive a score between [0, 1]. We define the score as the *label-preserving* (**LP**) factor of a transformation; a larger LP factor denotes a smaller impact on image aesthetics. As in Table II, *reflection* and *random*

TABLE II
THE SUBJECTIVE EVALUATION SURVEY ON THE AESTHETICS INFLUENCES OF VARIOUS TRANSFORMATIONS (s DENOTES A RANDOM NUMBER)

Transformation	Description	LP factor
Reflection	Flipping the image horizontally	0.99
Random scaling	Scale the image proportionally by $s \in$ [0.9, 1.1]	0.94
Small noise	Add a Gaussian noise $\in N(0, 5)$	0.87
Large noise	Add a Gaussian noise $\in N(0, 30)$	0.63
Alter RGB	Perturbed the intensities of the RGB channels [4]	0.10
Rotation	Randomly-parameterized affine transformation	0.26
Squeezing	Change the aspect ratio by $s \in$ [0.8, 1.2]	0.55

scaling receive the highest LR factors. The small noise seems to affect the aesthetics feelings negatively, but only marginally. All others are shown to significantly degrade human aesthetics perceptions. We therefore adopt reflection, random scaling, and small noise as our default data augmentation approaches.

VI. EXPERIMENT

A. Binary Rating Prediction

We implement our models based on the cuda-convnet package [4]. The ReLU nonlinearity as well as dropout is applied. Following RAPID [7], we evaluate DCM on the binary aesthetics rating task. We quantize images' mean ratings into binary values. Images with mean ratings smaller than $5-\delta$ are labeled as "low-quality", while those with mean ratings larger than $5+\delta$ are referred to as "high-quality". For the distribution prediction, we do not quantize the ratings.

The adjustment of learning rates in such a hierarchical model calls for special attentions. We first train the 14 parallel pathways, with the identical learning rates: η = 0.05 for unsupervised pre-training and 0.01 for supervised tuning, both of which are not annealed throughout training. We then train the high-level synthesis network on top of them and fine-tune the entire DCM. For the pathway part, its learning rate η' starts from 0.001; for the high-level part, the learning rate ρ starts from 0.01. When the training curve reaches a plateau, we first try dividing ρ by 10; and further try dividing ρ by 10 if the training/validation error still does not decrease.

Static Regularization v..s. Joint Tuning The RAPID model [7] also extracted attributes along different columns (pathways) and combine them. The pre-trained style classifier was then "frozen" and acted as a static network regularization. Out of curiosity, we also tried to fix our parallel pathways while training the high-level synthesis network, e.g., η' = 0. The resulting performance was verified to be inferior to that of joint tuning the entire DCM.

We compare DCM with the state-of-the-art RAPID model for binary aesthetics rating prediction. Benefiting from our fully-convolutional architecture, DCM has a much lower parameter capacity than RAPID that relies on fully-connected layers. Besides, we construct three baseline networks, all with exactly the same parameter capacity as DCM:

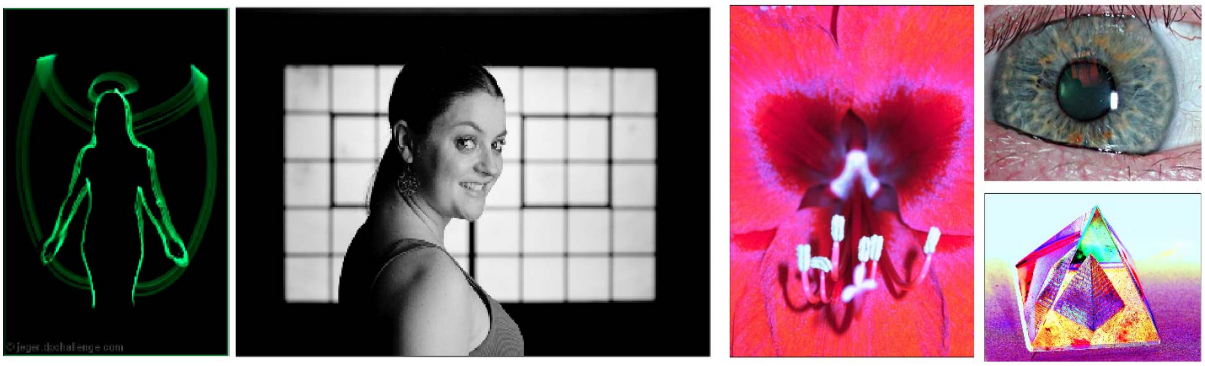
- **Baseline fully-convolutional network (BFCN)** first binds conv1 – conv3 layers of 14 pathways horizontally, constituting a three-layer fully convolutional network, each layer with 64 $\times$ 14 = 896 filter channels. Such a attribute learning part is trained in a unsupervised way, with style annotations utilized. It is then concatenated with the high-level synthesis network, to be jointly supervised-tuned.
- **DCM without parallel pathways (DCM-WP)** utilize style annotations in an entangled fashion. Its only difference with BFCN lies in that, the training of the attribute learning part is supervised by a composite label $\in R^{28\times 1}$, which binds 14 individual labels altogether.
- **DCM without data augmentations (DCM-WA)** denotes DCM without the three data augmentations applied (reflection, scaling, and small noise).

We train the above five models for the binary rating prediction, with both δ = 0 and δ = 1. The overall accuracies are compared in Table III. [3] It appears that BFCN performs significantly worse than others, due to the absence of the style attribute information. While RAPID, DCM-WP and DCM all utilize style annotations as the supervision, DCM outperforms the other two in both cases with remarkable margins. By comparing DCM-WP with DCM, we observe that the biologically-inspired parallel pathway architecture in DCM facilitates the learning. Such a specific architecture avoids overly large all-in-one models (such as DCM-WP), but instead have more effective, dedicated sub-models. In DCM, style annotations serve as powerful priors, to enforce DCM to focus on extracting features that are highly correlated to aesthetics judgements. The DCM is jointly tuned from end to end, which is different from RAPID whose style column only acts as a static regularization. We also notice a gain of nearly 3% of DCM over DCM-WA, which verifies the effectiveness of our proposed augmentation approaches.

In [8], a linear classifier was trained on fisher vectors computed from color and SIFT descriptors. Under the same aesthetic quality categorization setting, the baselines reported by [8] were 66.7% when σ = 0, and 67.0% when σ = 1, falling far behind both DCM and RAPID.

To qualitatively analyze the results, we display eight images correctly classified by DCM to be high-quality when δ = 0, in Fig. 3 (a), and eight correctly classified low-quality images in in Fig. 3 (b). The images ranked high in terms of aesthetics typically present salient foreground objects, low depth of field, proper composition, and color harmony. In contrast, low-quality images are at least defected in one aspect. For example, the

[3]The accuracies of RAPID are from the RDCNN results in Table 3 [7]



(a) (b)

Fig. 4. How contexts and emotions could alter the aesthetics judgment. (a) Incorrectly classified examples ($\delta = 0$) due to semantic contents; (b) High-variance examples (correctly predicted by DCM), which have nonconventional styles or subjects.

TABLE III
THE ACCURACY COMPARISON OF DIFFERENT METHODS FOR BINARY RATING PREDICTION.

	RAPID	BFCN	DCM-WP	DCM-WA	DCM
$\delta = 0$	74.46%	70.20%	73.54%	74.03%	76.80%
$\delta = 1$	73.70%	68.10%	72.23%	73.72%	76.04%

TABLE IV
THE AVERAGE KL DIVERGENCE COMPARISON FOR RATING DISTRIBUTION PREDICTION.

DCM	DCM-soft-D	DCM-KL-D
0.1743	0.2338	0.2052

top left image has no focused foreground object, while the bottom right one suffers from a messy layout. For the top right "girl" portrait in Fig 3 (b), we investigated its original comments on DPChallenge.com, and found that people rated it low because of the noticeable detail loss caused by noise reduction post-processing, as well as the unnatural "plastic-like" lights on her hair.

More interestingly, Fig. 4 (a) lists two **failure** examples of DCM. The left image in Fig. 4 (a) depicts a waving glowstick captured by time-lapse photography. The image itself has no appealing composition or colors, and is thus identified by DCM to be low-quality. However, the DPChallenge raters/commenters were amazed by the angel shape and rated it very favorably due to the creative idea. The right image, in contrast, is a high-quality portrait, on which DCM confidently agrees. However, it was associated with the "Rectangular" challenge topic on DPChallenge, and was rated low because this targeted theme was overshadowed by the woman. The failure examples manifest the tremendous inherent subjectivity and sensitivity of human aesthetics judgement.

B. Rating Distribution Prediction

To our best knowledge, among all state-of-the-art models working on latest large-scale datasets, DCM is the only one accounting for rating distribution prediction. We use the binary prediction DCM as the initialization, and re-train only the high-level synthesis network with the loss defined in Eqn. (1). We then compare the predicted distributions with the groundtruth of the AVA testing set. We also include two more DCM variants as baselines in this task:

- **DCM with the softmax loss for rating distribution vectors (DCM-soft-D)** makes the only architecture change by modifying the global average pooling of the high-level network to be 10-channel. Its output is compared to the raw rating distribution under the conventional softmax loss (i.e., cross entropy).
- **DCM with the KL loss for rating distribution vectors (DCM-KL-D)** replaces the softmax loss in DCM-soft-D, with the general KL loss (i.e., relative entropy) [32]. It remains to work with the raw rating distribution.

As compared in Table IV, KL-based loss function tends to perform better than the softmax function for this specific task. It is important to notice that DCM further reduces the KL divergence compared to DCM-KL-D. While the raw ratings can be noisy due to both the coarse rating grid and the limited rating number, we are able to obtain a more robust estimation of the underlying rating distribution, with the aid of the strong Gaussian prior from the AVA study [8].

Very notably, we observe that for more than 96% of the AVA testing images, the differences between their groundtruth mean values and estimates by DCM are less than 1. We further binarize the estimated and groundtruth mean values, to re-evaluate the results in the context of binary rating prediction. The overall accuracies are improved to 78.08% ($\delta = 0$), and 77.27% ($\delta = 1$). It verifies the benefits to jointly predict the means and standard deviations, built upon the AVA observation that they are correlated.

Fig. 4 (b) visualizes images that are correctly predicted by DCM to have large variances. It is intuitive that images with a high variance seem more likely to be edgy or subject to interpretation. Taking the top right image for example, the comments it received indicate that while many voters found the photo striking (e.g. "nice macro""good idea"), others found it rude (e.g. "it frightens me""too close for comfort").

VII. Conclusion

In this paper, we get inspired by the knowledge abstracted from the human visual perception and neuroaesthetics, and formulate the Deep Chatterjee's Machine (DCM). The biological inspired, task-specific architecture of DCM leads to superior performance, compared to other state-of-the-art models with the same or higher parameter capacity. Since it has been observed in Fig. 4 that emotions and contexts could alter the aesthetics judgments, we plan to take the two factors into account for a more comprehensive framework.

References

[1] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 291–300.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 288–301.

[3] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 419–426.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4792–4800.

[6] Z. Wang, J. Yang, H. Jin, E. Shechtman, A. Agarwala, J. Brandt, and T. S. Huang, "Deepfont: Identify your font from an image," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 451–459.

[7] 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*. ACM, 2014, pp. 457–466.

[8] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2408–2415.

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

[10] C. J. Cela-Conde, L. Agnati, J. P. Huston, F. Mora, and M. Nadal, "The neural foundations of aesthetic appreciation," *Progress in neurobiology*, vol. 94, no. 1, pp. 39–48, 2011.

[11] A. Chatterjee, "Neuroaesthetics: a coming of age story," *Journal of Cognitive Neuroscience*, vol. 23, no. 1, pp. 53–62, 2011.

[12] ——, "Prospects for a cognitive neuroscience of visual aesthetics," *Bulletin of Psychology and the Arts*, vol. 4, no. 2, pp. 55–60, 2003.

[13] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 271–280.

[14] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1657–1664.

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

[16] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1784–1791.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[18] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[19] Z. Dong, X. Shen, H. Li, and X. Tian, "Photo quality assessment with dcnn that understands image well," in *MultiMedia Modeling*. Springer, 2015, pp. 524–535.

[20] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *Signal Processing Magazine, IEEE*, vol. 28, no. 5, pp. 94–115, 2011.

[21] Z. Gao, S. Wang, and Q. Ji, "Multiple aesthetic attribute assessment by exploiting relations among aesthetic attributes," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 575–578.

[22] J. J. Nassi and E. M. Callaway, "Parallel processing strategies of the primate visual system," *Nature Reviews Neuroscience*, vol. 10, no. 5, pp. 360–372, 2009.

[23] J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar *et al.*, "Functional network organization of the human brain," *Neuron*, vol. 72, no. 4, pp. 665–678, 2011.

[24] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[25] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 11, pp. 1254–1259, 1998.

[26] J. K. Tsotsos and A. Rothenstein, "Computational models of visual attention," *Scholarpedia*, vol. 6, no. 1, p. 6201, 2011.

[27] G. Field and E. Chichilnisky, "Information processing in the primate retina: circuitry and coding," *Annu. Rev. Neurosci.*, vol. 30, pp. 1–30, 2007.

[28] G. A. Rousselet, S. J. Thorpe, and M. Fabre-Thorpe, "How parallel is visual processing in the ventral pathway?" *Trends in cognitive sciences*, vol. 8, no. 8, pp. 363–370, 2004.

[29] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2018–2025.

[30] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.

[31] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[32] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[33] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs the method of paired comparisons," *Biometrika*, vol. 39, no. 3-4, pp. 324–345, 1952.

[34] Z. Wang, Y. Yang, Z. Wang, S. Chang, J. Yang, and T. S. Huang, "Learning super-resolution jointly from external and internal examples," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4359–4371, 2015.