Rating Image Aesthetics using Deep Learning

Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James. Z. Wang

***Abstract*—This paper investigates unified feature learning and classifier training approaches for image aesthetics assessment. Existing methods built upon handcrafted or generic image features and developed machine learning and statistical modeling techniques utilizing training examples. We adopt a novel deep neural network approach to allow unified feature learning and classifier training to estimate image aesthetics. In particular, we develop a double-column deep convolutional neural network to support heterogeneous inputs, i.e., global and local views, in order to capture both global and local characteristics of images. In addition, we employ the style and semantic attributes of images to further boost the aesthetics categorization performance. Experimental results show that our approach produces significantly better results than the earlier reported results on the AVA dataset for both the generic image aesthetics and content-based image aesthetics. Moreover, we introduce a 1.5 million image dataset (IAD) for image aesthetics assessment and we further boost the performance on the AVA test set by training the proposed deep neural networks on the IAD dataset.**

***Index Terms*—Automatic feature learning, deep neural networks, image aesthetics**

I. INTRODUCTION

Automated assessment of image aesthetics is a significant research problem due to its potential applications in areas where visual experience is involved, such as image retrieval, image editing, design, and human computer interaction. Assessing aesthetics of images is challenging for computers because aesthetics may be rated differently by different people, and an optimal computational representation of aesthetics is not obvious. More importantly, the difficulty lies in designing a proper image representation (features) to map the perception of images to their aesthetics ratings.

Among the initial attempts, image aesthetics was represented by discrete values, and the problem of assessing image aesthetics was formulated as either a classification or regression problems [1], [2]. In the past decade, many visual features have been explored under this formulation (handcrafted features), ranging from low-level image statistics, such as edge distributions and color histograms, to high-level photographic rules, such as the rule of thirds and golden ratio [1], [2], [3], [4], [5], [6], [7], [8].

These aesthetics-relevant features are often inspired and designed based on intuition in photography or psychology literature; however, they share some essential limitations. For example, some aesthetics-relevant attributes may be unexplored and thus poorly defined as objective criteria. Meanwhile, most photographic or psychological rules are descriptive. The computed features are approximations of those rules, then, are limited to approximations. To overcome such limitations, image features commonly used for image classification or image retrieval (generic features) were applied to image aesthetics [9], [10], [11], such as SIFT and Fisher Vector [12], [9]. Whereas generic image features have shown better performance in [9] than handcrafted aesthetics features have, they may not provide an optimal representation for aesthetics-related problems due to their generic nature.

We are motivated by the feature learning power of deep convolutional neural networks [13], where feature learning is unified with classifier training using RGB images, and we propose to learn effective aesthetics features using convolutional neural networks. However, applying classic architecture to our task is not straightforward. Image aesthetics depends on a combination of local cues (e.g., sharpness and noise levels) and global visual cues (e.g., the rule of thirds). To learn aesthetics-relevant representations of an image, we generate two heterogeneous inputs to represent its global cues and local cues respectively, as shown in Figure 1. Meanwhile, we develop a double-column neural network architecture that takes parallel inputs from the two columns to support network training on heterogeneous inputs, which extends the method in [13]. In the proposed double-column neural network, one column takes a global view of the image and the other column takes a local view of the image. The two columns are aggregated after some layers of transformations and mapped to the label layer.

We apply the proposed double-column neural network approach to the generic image aesthetics problem and propose a network adaptation approach for content-based image aesthetics. We also propose a regularized neural network approach by exploring related attributes such as style and semantic attributes associated with images. We show that our approaches achieve state-of-the-art results on the recently-released AVA dataset and further improve the performance by introducing a 1.5 million dataset (IAD) and performing network training using the IAD dataset.

A. Related Work

In this section, we review the handcraft and generic image features that have been explored for image aesthetics. We also review the successful applications of deep convolutional neural networks in recent studies.

Common visual cues such as color [1], [7], [8], texture [1], [2], composition [3], [5], [6], and content [5], [6] have been examined in earlier visual aesthetics assessment research. Color features typically include lightness, colorfulness, color harmony, and color distribution [1], [7], [8]. Texture descriptors

This paper extends: X. Lu, Z. Lin, H.L. Jin, J.C. Yang, and J.Z. Wang. "RAPID: Rating Pictorial Aesthetics Using Deep Learning," in *ACM International Conference on Multimedia*, pp. 457-466, 2014.

X. Lu and J. Wang are with College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, 16802, USA. E-mail: {xinlu,jwang}@psu.edu

Z. Lin, H.L. Jin, and J.C. Yang are with Adobe Research, Adobe Systems Inc., San Jose, CA, 95110, USA.
E-mail: {zlin,hljin,jiayang}@adobe.com

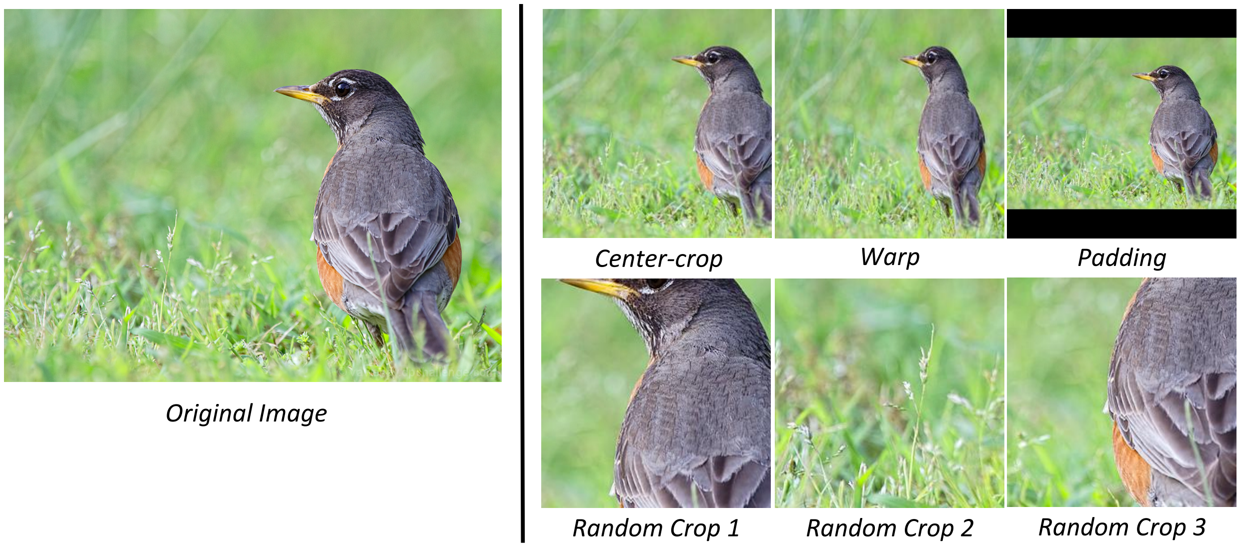


Fig. 1. Global views and local views of an image. Global views are represented by normalized inputs: center-crop, warp, and padding (shown in the top row). Local views are represented by randomly-cropped inputs from the original high-resolution image (examples shown in the bottom row).

cover wavelet features [1], edge distributions, blur descriptors, and shallow depth-of-field features [2]. Composition features range from the rule of thirds, size and aspect ratio [5] to foreground and background composition [3], [5], [6]. Meanwhile, the content of images has been studied by taking account of people and portrait descriptors [5], [6], scene descriptors [6], and generic image features [9], [10], [11].

Despite the success of handcrafted and generic features for analyzing image aesthetics problems, unifying the automatic feature learning and classifier training using deep neural networks has shown promising performance in various applications [13], [16], [17], [18]. In particular, convolutional neural network (CNN) [19] is one of the most powerful deep learning architectures in vision problems; other deep learning architectures include Deep Belief Net [20] and Restricted Boltzmann Machine [21]. In [13], Krizhevsky *et al.* significantly improved the image classification performance on the ImageNet benchmark using CNN, along with dropout and normalization techniques. In [18], Sermanet *et al.* achieved the best performance compared with other reported results on all major pedestrian detection datasets. In [16], Ciresan *et al.* reached a near-human performance on the MNIST[1] dataset. The effectiveness of extracted CNN features has also been demonstrated in image style classification [22] and image popularity estimation [25].

Few studies have investigated automatic feature learning for image aesthetics prediction, as designing handcrafted features has long been regarded as an appropriate method in predicting image aesthetics. The emergence of the AVA dataset, containing $250,000$ images with aesthetics ratings, makes it possible to learn features automatically and assess image aesthetics using deep learning.

In this study, we systematically evaluate deep neural networks on the problem of image aesthetics assessment. In particular, we develop a double-column CNN to capture image aesthetics-relevant features from two heterogeneous input sources and improve the image aesthetics prediction accuracy given new images. The proposed architecture is different from recent efforts on multi-column neural networks [16], [24]. In [24], Agostinelli *et al.* extended stacked sparse autoencoder to a multi-column version, computed the optimal column weights, and then applied the model to image denoising. In [16], Ciresan *et al.* averaged the output of several columns, where each column was associated with training input produced by different standard preprocessing methods. Unlike [16], the two columns in our architecture are jointly trained using two input sources: One column takes a global view as the input, and the other column takes a local view as the input. Such an approach allows us to capture both global and local visual information of images.

[1] http://yann.lecun.com/exdb/mnist/

B. Contributions

Our main contributions are as follows.

- We systematically evaluate the single-column deep convolutional neural network approach using different types of input modalities to predict image aesthetics.
- We develop a novel double-column deep convolutional neural network architecture to capture both global and local information of images.
- We develop a network adaptation-based approach to perform content-based aesthetic categorization.
- We develop a regularized double-column deep convolutional neural network to further improve aesthetic categorization using style attributes and semantic attributes.
- We introduce a 1.5 million dataset (IAD) and further improve the aesthetics categorization accuracy on the AVA test set.

II. The Approach

Photographers' visual preferences are often indicated through patterns in aesthetically-pleasing photographs, where composition [26] and visual balance [27] are two major factors [28]. Such factors are reflected in both global and local views of images. For instance, we present the global views of

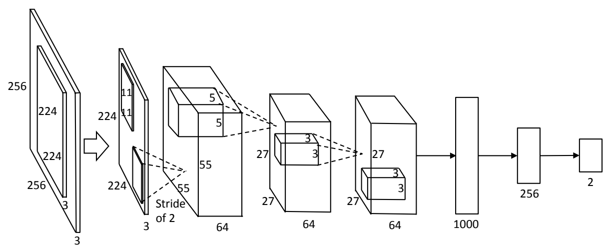


Fig. 2. Single-column convolutional neural network for image aesthetics assessment. The network architecture consists of four convolutional and two fully-connected layers. The max-pooling and normalization layers are following the first and second convolutional layers. To avoid overfitting, the input patch ($224 \times 224 \times 3$) is randomly cropped from the normalized input ($256 \times 256 \times 3$) as performed in [13].

an image in the top row and the local views in the bottom row (as shown in Figure 1). Commonly used composition rules in photography include the rule of thirds, diagonal lines, and the golden ratio [29]; and visual balance is closely connected with position, form, size, tone, color, brightness, contrast, and proximity to the fulcrum [27]. Generating accurate computational representation of these patterns is highly challenging because those rules are usually descriptive and vague in terms of definition. This issue motivates us to explore deep network training approaches that can automatically learn aesthetics-relevant features and perform aesthetics prediction.

Applying CNN to the problem of image aesthetics is not straightforward. The CNN is commonly trained on normalized training examples with fixed size and aspect ratio. However, in image aesthetics, normalizing images may lose important information because the visual perception of aesthetics is influenced by both the global view and local details. To address this difficulty, we propose using heterogeneous representations of an image as training examples. In doing so, we expect the deep networks to be able to capture both global and local views jointly for image aesthetics assessment.

In the following sections, we first review single-column CNN (SCNN) training and evaluate the performance of SCNN using different outputs. We then present the proposed double-column CNN (DCNN) architecture and the design rationale. We also introduce the network adaptation for content-based image aesthetics. Finally, we study how to leverage external attributes to help image aesthetics assessment. We present the regularized double-column network (RDCNN) architecture, to perform network training for image aesthetics using style and semantic attributes, respectively.

A. Single-column Convolutional Neural Network

Deep convolutional neural network [13] is commonly trained using inputs of fixed aspect ratio and size; however, images could be of arbitrary size and aspect ratio. To normalize input images, a conventional approach is to isotropically resize original images by normalizing their shorter sides to a fixed length s, and crop the center patch as the input [13]. We refer to this approach as center-crop (g_c). In addition to g_c, we attempted two other transformations to normalize images, i.e., warp (g_w) and padding (g_p), in order to represent the global view (I_g) of an image I. g_w anisotropically resizes (or warps) the original image into a normalized input with a fixed size $s \times s \times 3$. g_p resizes the original image by normalizing the longer side of the image to a fixed length s and padding border pixels with zeros to generate a normalized input of a fixed size $s \times s \times 3$. For each image I and each type of transformation, we generate an $s \times s \times 3$ input I_g^j with the transformation g_j, where $j \in \{c, w, p\}$. Because normalizing inputs may cause harmful information loss (i.e., the high-resolution local views) for aesthetics assessment, we randomly sampled fixed size (at $s \times s \times 3$) crops with the transformation l_r from original high-resolution images. Here we denote by g the global transformations and l the local transformations. This results in a collection of normalized inputs $\{I_l^r\}$ (r refers to an index of normalized inputs in the collection), which preserve the local details on the original high-resolution image. We took these normalized inputs $I_t \in \{I_g^c, I_g^w, I_g^p, I_l^r\}$ for SCNN training[2].

We show examples of the four transformations, g_w, g_c, g_p, and l_r, in Figure 1. In the top row, we present the global views of an image depicted by g_c, g_w, and g_p. It is clear that I_g^w and I_g^p maintain the relative spatial layout among elements in the original image while the I_g^c does not. In the bottom row, we show that the local views of an original image are represented by randomly-cropped patches $\{I_l^r\}$, which describe the local details in the original high-resolution image.

We present the architecture of the SCNN used for image aesthetics in Figure 2. The network architecture consists of four convolutional and two fully-connected layers. The max-pooling and normalization layers follow the first and second convolutional layers. To avoid overfitting, the input patch ($224 \times 224 \times 3$) is randomly cropped from the normalized input ($256 \times 256 \times 3$), as performed in [13].

For the input I_p of the i-th image, we denote by $\mathbf{x}_i$ the feature representation extracted from the fc256 layer, i.e., the outcome of the convolutional layers and the fc1000 layers, and $y_i \in \mathcal{C}$ the label. We maximize the following log likelihood function to train the last layer:

$$l(\mathbf{W}) = \sum_{i=1}^{N} \sum_{c \in \mathcal{C}} \mathbb{I}(y_i = c) \log p(y_i = c \mid \mathbf{x}_i, \mathbf{w}_c) \quad , \qquad (1)$$

where N is the number of images, $\mathbf{W} = \{\mathbf{w}_c\}_{c \in \mathcal{C}}$ is the set of model parameters, and $\mathbb{I}(x) = 1$ iff x is true and vice versa. The probability $p(y_i = c \mid \mathbf{x}_i, \mathbf{w}_c)$ is expressed as

$$p(y_i = c \mid \mathbf{x}_i, \mathbf{w}_c) = \frac{\exp\left(\mathbf{w}_c^T \mathbf{x}_i\right)}{\sum_{c' \in \mathcal{C}} \exp\left(\mathbf{w}_{c'}^T \mathbf{x}_i\right)} \quad . \qquad (2)$$

In our experiments, image aesthetics assessment is formulated as a two-class classification problem, where each input is associated with an aesthetic label $c \in \mathcal{C} = \{0, 1\}$. The image style classification to be discussed in Section II-C is formulated as a multi-class classification task.

The general guideline that we have found to train a deep network is first to allow sufficient learning capacity by using a sufficient number of neurons. Meanwhile, we adjust the number of convolutional layers and the fully-connected layers to support automatic feature learning and classifier training.

[2] In our experiments, we set s to be 256, and the size of I_t is $256 \times 256 \times 3$.

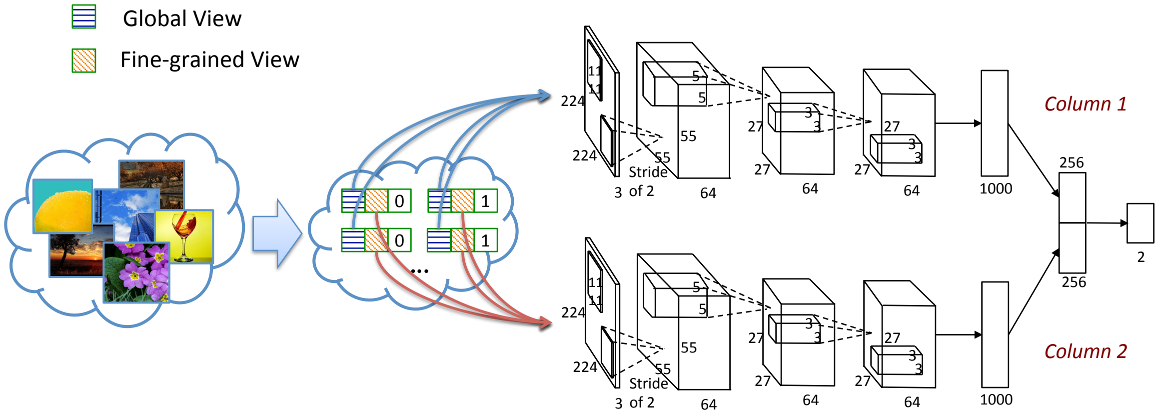


Fig. 3. Double-column convolutional neural network for aesthetic quality categorization. Each training image is represented by both global and local views, and is associated with its aesthetic quality label: 0 refers to a low quality image and 1 refers to a high quality image. Networks in different columns are independent in convolutional and the first two fully-connected layers. All parameters in DCNN are jointly trained trained.

We then evaluate networks trained using different numbers of convolutional and fully-connected layers, and with or without normalization layers. Finally, we conduct empirical evaluations on candidate architectures under the same experimental settings. An empirically optimal architecture for a specific task can then be selected. In our experiments, we list candidate architectures, and we determine an empirically optimal architecture for our task by conducting experiments on candidate architectures using the same experimental settings and picking the one that has achieved the best performance.

Using the selected network architecture, we trained and evaluated SCNN with four types of inputs (I_g^c, I_g^w, I_g^p, I_l^r) on the AVA dataset [10]. In training, we adopted dropout and shuffled the training data in each epoch to alleviate overfitting. Interestingly, we found that l_r might be an effective data augmentation approach. Because I_l^r is generated by random cropping, one image is actually represented by different random patches in different epochs.

Given a test image, we computed its normalized input and generated the input patch. We then computed the probability of the input patch being assigned to each aesthetics category. We repeated the process 50 times and averaged those results to identify the class with the highest probability as the prediction.

B. Double-column Convolutional Neural Network

Transforming an input image to a certain normalized input (g_c, g_w, g_p, or l_r) may result in information loss of either the global view or local details. This potential loss motivates us to explore the network training approach to support heterogeneous inputs. To achieve this goal, we propose a novel double-column convolutional neural network (DCNN), allowing network training using two inputs extracted from different spatial scales of one image. We present the DCNN architecture in Figure 3, where the two columns are independent in the convolutional and the first two fully-connected layers. Then, the two output vectors of the fc256 layers are concatenated and mapped to the label layer. The interaction between the two columns happens at a later fully-convolutional layer to allow sufficient flexibility in feature learning from heterogeneous inputs. In DCNN network training, all parameters in DCNN are jointly trained, which enables the network to judge image aesthetics while simultaneously considering both the global and local views of an image. Specifically, the error is back propagated to the networks in each column respectively with stochastic gradient descent.

As shown, the proposed DCNN architecture could easily be expanded to multiple columns and support multiple types of normalized inputs as training examples. Even more flexible, the DCNN allows different architectures and initializations in individual networks prior to their interaction, which usually happens at a later fully-connected layer. Such design facilitates parameter learning, especially in the case of multiple-column architectures. To predict the aesthetics value of a new image, we follow the same procedure as we evaluate the SCNN for image aesthetics assessment.

Given an image of a semantic category, is there a better solution to estimate its image aesthetics besides applying the trained DCNN on images of any semantic category? A most straightforward approach is to collect images of a specific semantic category that are associated with aesthetic labels and train DCNN on that image collection. Unfortunately, collecting such datasets and training individual DCNN for each of the semantic category is time-consuming. We expect to develop a generic network that can be reused for content-based image aesthetics, where the number of images in each semantic category is not large. To fit that purpose, we build upon the DCNN network structure, and propose a network adaptation strategy to approach content-based image aesthetics. Given a DCNN network trained on a large collection of images with arbitrary content, we adaptively update the DCNN network parameters in a few training epochs using a small collection of images in a specific semantic category. The advantages are that images of all semantic categories are used for training as well as for simplifying the data collection process. Importantly, the time used for network adaptation is much shorter than training a network from scratch. We demonstrate the performance of network adaptation in the experimental section.

C. Learning and Categorization with Style and Semantic Attributes

Images are divided into aesthetics categories by quantizing their aesthetics values. Limited categories (i.e., high and low

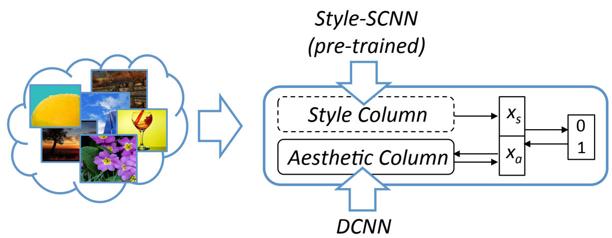


Fig. 4. Regularized double-column convolutional neural network (RDCNN). The style attributes are used as extra attributes to facilitate network training for image aesthetics. The style attributes $\mathbf{x}_s$ are produced by pre-trained Style-SCNN, leveraging which we regularize the aesthetics column network training. The dashed line indicates that the parameters of the style column is fixed during RDCNN training. In backpropagation, only the parameters in the aesthetic column are fine-tuned, and the parameter fine-tuning process is supervised by the aesthetic label. We used style attributes as an example to show how extra attributes are leveraged in RDCNN training. The same procedures are applied to semantic attributes.

aesthetics in this work) result in large intra-class variation in terms of image content. This limitation makes the network training difficult for the classification tasks because the label may not sufficiently guide the network training. This motivates us to exploit extra attributes of images as a complement of aesthetics labels in order to facilitate network training. Considering style and semantic attributes are highly relevant to perceived aesthetics [10], we propose to use those two types of attributes in this work as extra guidance for automatic feature learning. We first formulate the aesthetic categorization problem with style attributes, and we then apply the same approach to categorize image aesthetics using semantic attributes.

We have formulated the problem in two ways to leverage the extra attributes. The first solution is to borrow the idea from multi-task learning [31], where feature representation and the classification error minimization are constructed in a joint manner for both labels. Assuming we have aesthetics quality labels $\{y_{ai}\}$ and style labels $\{y_{si}\}$ for all training images, the problem could be formulated as:

$$\begin{aligned}\max_{\mathbf{X},\mathbf{W}_a,\mathbf{W}_s} \sum_{i=1}^{N} (\sum_{c\in\mathcal{C}_A} \mathbb{I}(y_{ai}=c)\log p(y_{ai} \mid \mathbf{x}_i, \mathbf{w}_{ac}) + \\ \sum_{c\in\mathcal{C}_S} \mathbb{I}(y_{si}=c)\log p(y_{si} \mid \mathbf{x}_i, \mathbf{w}_{sc})) \quad ,\end{aligned} \tag{3}$$

where $\mathbf{X}$ is the features of all training images, $\mathcal{C}_A$ is the label set for aesthetic quality, $\mathcal{C}_S$ is the label set for style (or semantics), and $\mathbf{W}_a = \{\mathbf{w}_{ac}\}_{c\in\mathcal{C}_A}$ and $\mathbf{W}_s = \{\mathbf{w}_{sc}\}_{c\in\mathcal{C}_S}$ are the model parameters. Nevertheless, associating all images with style attributes is not easy, and we are lack of such datasets where images are associated with both aesthetics and style labels. The AVA dataset only contains $14,000$ images that have both aesthetics and style labels among the $230,000$ training images. Therefore, we are unable to follow the multi-task learning formulation due to the missing attributes in the training dataset.

Alternatively, we take ideas from inductive transfer learning [30], where we minimize the classification error with one label, while we construct feature representations with both labels. We first train a style classifier using the subset of images that are associated with style labels, and then extract style attributes for all training images. Using style attributes, we train regularized deep networks for image aesthetics assessments.

We learn style attributes by SCNN (introduced in Section II-A) using images associated with style labels in the AVA dataset. We denote it by Style-SCNN. We have also tried to use DCNN for style classification. Due to the limited number of training data $(11,000)$, the warped column did not contribute much to boost the style classification performance, so we skipped it. We present the architecture of Style-SCNN in Figure 2. Due to the smaller number of training examples, we reduced the number of filters in the first and fourth convolutional layers by half compared to the number of filters used in aesthetics network training. The style attributes are extracted as the output of the fc256 layer in the Style-SCNN.

To train deep networks for image aesthetics using extra attributes (such as styles and semantics), we propose a regularized double-column convolutional neural network (RDCNN). We show a RDCNN example using style attributes in Figure 4. As shown, two normalized inputs of the aesthetics column are I_g^w and I_l^r, same as in DCNN (Section II-B). The input of the style column is I_l^r. The training of RDCNN is done by solving the following optimization problem:

$$\max_{\mathbf{X}_a,\mathbf{W}_a} \sum_{i=1}^{N} \sum_{c=1\in\mathcal{C}_a} \mathbb{I}(y_{ai}=c)\log p(y_{ai} \mid \mathbf{x}_{ai}, \mathbf{x}_{si}, \mathbf{w}_{ac}) \quad , \tag{4}$$

where $\mathbf{x}_{si}$ is the style attributes of the i-th training image, and $\mathbf{x}_{ai}$ is the feature to be learned. In particular, the maximization does not involve style attributes $\mathbf{x}_s$, which means that we only fine-tuned the parameters in the aesthetic column in backpropagation and that the learning process is supervised by the aesthetic label. The parameters of the style column are fixed, and the style attributes $\mathbf{x_{is}}$ essentially serve as a regularizer to train the aesthetic column for image aesthetics assessment.

Similar to style attributes, semantic attributes of an image may also share high correlations with image aesthetics. For instance, images of a cute baby and images of gardens or beautiful scenes may be aesthetically more appealing. This motivates us to use semantic attributes to regularize the aesthetics network training using the proposed RDCNN. As limited semantic tags are associated with images in the AVA dataset, we took pre-trained ImageNet model as the pre-trained column and conducted regularized DCNN training using the proposed RDCNN approach. The parameters of the ImageNet column are fixed and the semantic attributes regularized the network training in the aesthetics columns.

III. Experimental Results

We evaluated the proposed methods for image aesthetics on the AVA[3] dataset [10] and the IAD dataset. On the AVA dataset, we divided training images into two categories, i.e., low-quality and high-quality images, according to criteria

[3]The AVA dataset includes $250,000$ images, each of which is associated with an aesthetics score, averaged by about more than 200 ratings. The scale of the aesthetics score is from 1 to 10. We took the same experimental settings as in [10]. Particularly, we took the same division of training and testing data as in [10], i.e., $230,000$ images for training and $20,000$ for testing.

TABLE I
ACCURACY FOR DIFFERENT SCNN ARCHITECTURES

	conv1 (64)	pool1	rnorm1	conv2 (64)	pool2	rnorm2	conv3 (64)	conv4 (64)	conv5 (64)	conv6 (64)	fc1K	fc256	fc2	Accuracy
Arch 1	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$			$\surd$	$\surd$	$\surd$	$\mathbf{71.20\%}$
Arch 2	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$					$\surd$	$\surd$	$\surd$	60.25%
Arch 3	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$				$\surd$	$\surd$	$\surd$	62.68%
Arch 4	$\surd$	$\surd$		$\surd$	$\surd$		$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	65.14%
Arch 5	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$				$\surd$	$\surd$	70.52%
Arch 6	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$		$\surd$	$\surd$	$\surd$	62.49%
Arch 7	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	70.93%

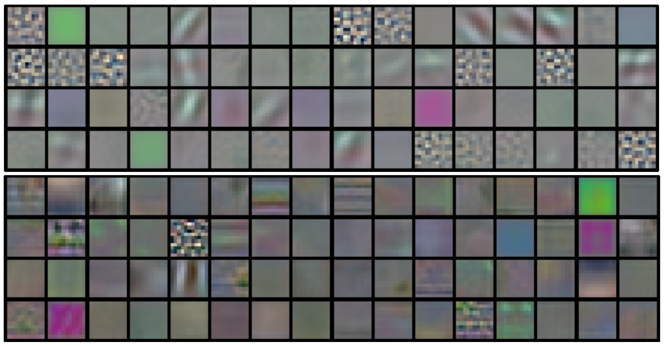


Fig. 5. Filter visualization of DCNN for image aesthetics. In particular, 128 convolutional kernels of the size $11 \times 11 \times 3$ learned by the first convolutional layer. The first 64 are from the local view column (with the input I_l^r) and the last 64 are from the global view column (with the input I_g^w).

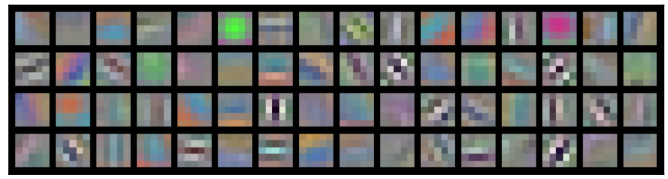


Fig. 6. Filter visualization of CNN for image classification on CIFAR dataset. 64 convolutional kernels of the size $5 \times 5 \times 3$ learned by the first convolutional layer.

TABLE II
ACCURACY OF AESTHETIC CATEGORIZATION WITH DIFFERENT INPUTS

δ	I_l^r	I_g^w	I_g^c	I_g^p
0	71.20%	67.79%	65.48%	60.43%
1	68.63%	68.11%	69.67%	70.50%

presented in [10], and we learned style attributes using the AVA style dataset[4].

We first present the performance of SCNN using different network architectures and taking I_l^r as the input. We select the best performing architecture and evaluate SCNN using different types of inputs. Next, we present image aesthetics prediction results produced by DCNN, and qualitatively analyzed the advantage of the double-column architecture over a single-column one. We evaluated the network adaptation approach for content-based image aesthetics on the eight representative image categories [10] (portrait, animal, stilllife, fooddrink, architecture, floral, cityscape, and landscape), and compared the network adaptation results (including both the SCNN and DCNN) with the state-of-the-art aesthetics accuracy in each of the eight categories. Further, we show the accuracy of trained style classifier and aesthetic categorization results generated by RDCNN with style attributes or semantic attributes incorporated. Moreover, we introduce a 1.5 million image dataset (IAD) with aesthetics scores, including images derived from DPChallenge[5] and PHOTO.NET[6]. We further boost the aesthetics assessment accuracy, presented in Section III-F, by training SCNN and DCNN on the IAD dataset. Finally, we discuss the computational efficiency of the proposed approaches on the AVA dataset and the IAD dataset.

[4]The AVA style dataset includes $11,000$ images for training and $2,500$ images for testing. Each of the images in the training set is associated with one of the 14 style labels, i.e., complementary colors, duotones, HDR, image grain, light on white, long exposure, macro, motion blur, negative images, rule of thirds, shallow DOF, silhouettes, soft focus, and vanishing point. Each of images in the test dataset is associated with one or multiple style labels as the ground truth.

[5]http://www.dpchallenge.com

[6]http://photo.net

A. SCNN Results

We first examine the performance of SCNN on different network architectures and present overall accuracy of image aesthetics using seven different architectures listed in Table I. The selected layer for each architecture is labeled with a check mark. To fairly compare the performance of network architectures, we took the same normalized input, I_l^r, as training examples, and we let $\delta = 0$. As shown in the Table, the highest accuracy was achieved by the Arch 1. We thus fixed the network architecture to Arch 1 in the following experiments.

We evaluate the performance of SCNN with various normalized inputs as training examples, i.e., I_g^c, I_g^w, I_g^p, and I_l^r. We trained deep networks with both $\delta = 0$ and $\delta = 1$ for each input type, and presented the overall accuracy in Table II. We observed from the Table that the best performance was achieved by I_l^r, which indicates that l_r is an effective data augmentation strategy to capture the fine-grained details pertinent to image aesthetics. We also noticed that I_g^w produces the highest accuracy among the three inputs for capturing the global view of images. We present the best performance of SCNN using Arch 1 and I_l^r as the training input in Table III. As shown, our performance is better than the previous study [10] for both $\delta = 0$ and $\delta = 1$.

B. DCNN Results

As we have shown that Arch 1 performs the best among all attempted architectures in Section III-A. In DCNN training and testing, we adopted the SCNN architecture Arch 1 for both columns. We took three inputs combinations to train the proposed double-column network, i.e., I_l^r and I_g^c, I_l^r and I_g^p, and I_l^r and I_g^w. Let $\delta = 0$, we empirically evaluated results achieved by the three variations. The combination of I_l^r and I_g^c achieves 71.8% accuracy, and I_l^r and I_g^p achieves 72.27% accuracy. The combination of I_l^r and I_g^w performs the best among the three, achieving 73.25% accuracy. Thus, we used the two inputs of I_l^r and I_g^w in DCNN training and evaluation.

TABLE III
ACCURACY OF AESTHETIC QUALITY CATEGORIZATION FOR DIFFERENT METHODS

δ	[10]	SCNN	AVG_SCNN	DCNN	RDCNN_style	RDCNN_semantic
0	66.7%	71.20%	69.91%	73.25%	74.46%	75.42%
1	67%	68.63%	71.26%	73.05%	73.70%	74.2%

(a) Images ranked the highest in aesthetics by DCNN

(b) Images ranked the lowest in aesthetics by DCNN

Fig. 7. Images ranked the highest and the lowest in aesthetics generated by DCNN. Differences between low-aesthetic images and high-aesthetic images heavily lie in the amount of textures and complexity of the entire image.

In Figure 5, we visualize the filters of the first convolutional layer in the trained DCNN. The first 64 filters are from the local column (using the input I_l^r), and the last 64 filters are from the global column (using the input I_g^w). For comparison, we showed filters trained in the object recognition task on CIFAR dataset[7] in Figure 6. Interestingly, we found that the filters learned with image aesthetic labels are free from radical intensity changes and look smoother and cleaner. Such observation indicates that differences between low-aesthetic and high-aesthetic cues primarily lie in the amount of texture

[7]http://www.cs.toronto.edu/~kriz/cifar.html

Fig. 8. Test images correctly classified by DCNN but misclassified by SCNN. The first row shows the images that are misclassified by SCNN with the input I_l^r. The second row shows the images that are misclassified by SCNN with the input I_g^w. The label on each image indicates the ground-truth aesthetic quality.

and complexity of the entire image. Such intuitions could also be observed from example images, presented in Figure 7. In general, the images ranked the highest in aesthetics are smoother than those ranked the lowest. This finding substantiates the significance of simplicity and complexity features recently developed for analyzing perceived emotions [32].

To further show the power of DCNN, we quantitatively compared its performance with that of the SCNN and [10]. We show in Table III that DCNN outperforms SCNN for both $\delta = 0$ and $\delta = 1$, and significantly outperforms the earlier study. We further demonstrate the effectiveness of joint training strategy adopted in DCNN by comparing DCNN with AVG_SCNN, which averaged the two SCNN results taking I_g^w and I_l^r as inputs. We present the comparisons in Table III, and the results show that DCNN performs better than the AVG_SCNN with both $\delta = 0$ and $\delta = 1$.

To analyze the advantage of the double-column architecture, we visualize test images correctly classified by DCNN and misclassified by SCNN. Examples are presented in Figure 8, where images in the first row are misclassified by SCNN with the input I_l^r, and images in the second row are misclassified with the input I_g^w. The label annotated on each image indicates the ground-truth aesthetic quality. We found that images misclassified by SCNN with the input I_l^r mostly dominated by an object, which is because the input I_l^r fails to consider the global information in an image. Similarly, images misclassified by SCNN with the input I_g^w usually contain fine-grained details in their local views. The result implies that both global view and fine-grained details help improve the aesthetics prediction accuracy as long as the information is properly leveraged.

As discussed in Section II-B, a natural extension of DCNN is to use multiple columns in CNN training, such as a quad-column CNN. Let $\delta = 0$, we attempted to use the four inputs I_l^r, I_g^c, I_g^p, and I_g^w to train a quad-column CNN. Compared with the DCNN, training a quad-column network architecture requires GPU with a larger memory. We adopted the SCNN architecture, the Arch 1, for all the four columns, and it turns out that a GPU with 5G memory (such as Nvidia Tesla M2070/M2090 GPU) is no longer applicable for network training with a mini-batch size of 128. Meanwhile, optimizing parameters in quad-column CNN is more difficult than a double-column CNN because an individual column may require a different learning rate and training duration. In practice, we initialized each of the four columns with SCNN trained using I_l^r, I_g^c, I_g^p, and I_g^w, respectively. We then fine-tuned the last layer of the quad-column neural network. Compared with DCNN, a quad-column CNN achieved a slightly higher accuracy of 73.38%.

C. Content-based Image Aesthetics

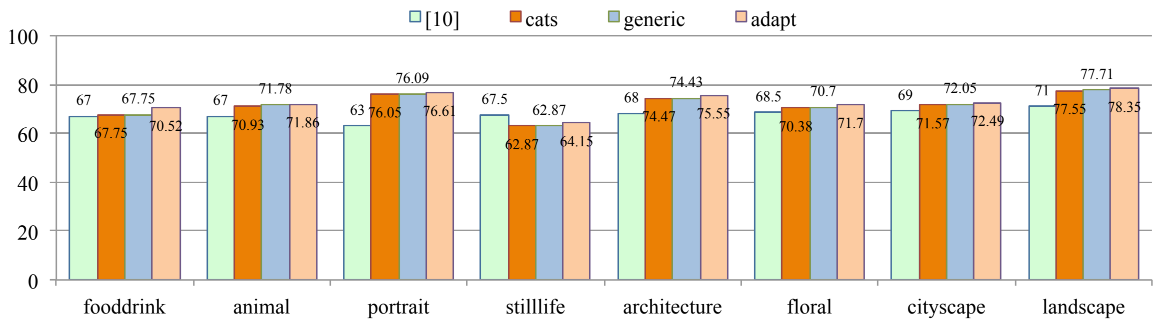
To demonstrate the effectiveness of network adaptation for content-based image aesthetics, we took the eight most popular semantic tags as used in [10]. We used the same training and testing image collection with [10], roughly $2.5K$ for training and $2.5K$ for testing in each of the categories[8].

In each of the eight categories, we systematically compared the proposed network adaptation approach (denoted by "adapt") built upon the SCNN (with the input I_l^r and I_g^w) and the DCNN with two baseline approaches ("cats" and "generic") and a state-of-the-art approach [10][9]. "cats" refers to the approach that trains the network using merely the categorized images (i.e., roughly $2.5K$ in each category), and "generic" refers to the approach that trains the network using the AVA training set (i.e., including images of arbitrary semantic categories). As presented in Figure 9 (a), (b), and (c), the proposed network training approach significantly outperforms the state of the art [10] (except the category of stilllife). In particular, the "generic" produces higher accuracy in general than the "cats" for SCNN with the input I_g^w and the DCNN, and "general" performs similar with "cats" for SCNN with the input I_l^r. This indicates the effectiveness of g_r. For the SCNN with both the inputs and the DCNN, "adapt" show better performance in most of the categories than "cats" and "generic".

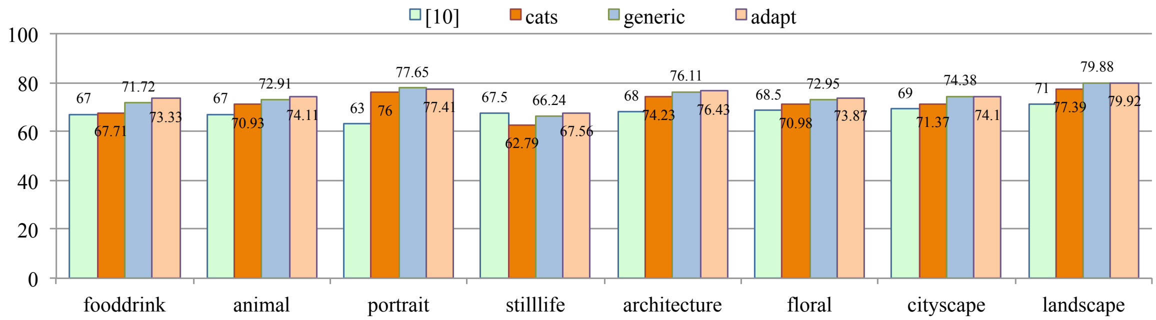
We also observed that the SCNN with the input I_g^w performs better than the SCNN with the input I_l^r, as shown in Figure 9. This result indicates that once an image is associated with an obvious semantic meaning, then the global view is more important than the local view in terms of assessing image aesthetics. Moreover, for both the "generic" and the "adapt",

[8]Few images in the AVA dataset have been removed from the Website, so we might have slightly smaller number of test images compared with [10]. Specifically, the number of test images in each of the eight categories is: portrait(2488), animal(2484), stilllife(2491), fooddrink(2493), architecture(2495), floral(2495), cityscape(2494), and landscape(2490).

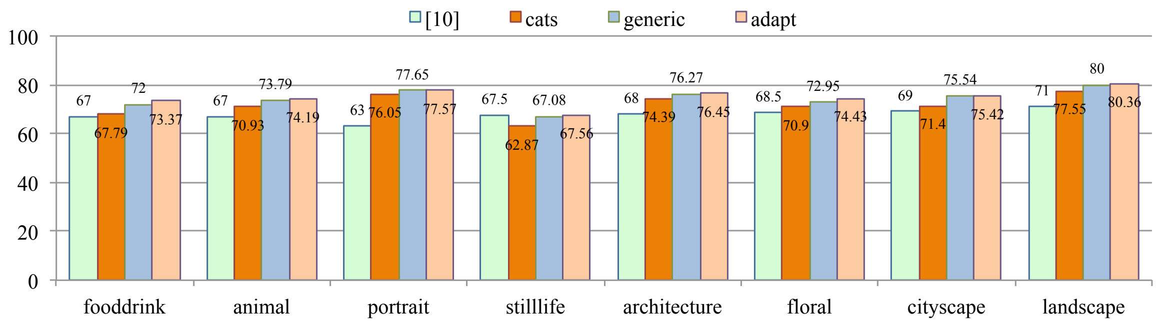
[9]We refer to the best performance of content-based image aesthetics in [10].



(a) SCNN with the input I_g^r



(b) SCNN with the input I_g^w



(c) DCNN

Fig. 9. Classification accuracy of image aesthetics in the eight semantic categories.

Fig. 10. Image examples in the stilllife category. Left: Images in the stilllife category. Right: Images with artistic styles in the stilllife category.

the DCNN outperforms the SCNN with the input I_g^w and I_l^r, which indicates that both the global view and the local view contribute to the aesthetic quality categorization of content-specific images. The DCNN does not show improvement in the "cats" due to the limited number of training examples. In combination with image classification [13], this content-based method can produce overall aesthetics predictions given uncategorized images.

We investigated the reason why the performance of SCNN is worse than [10] in the "stilllife" category, while in the

TABLE IV
ACCURACY FOR DIFFERENT NETWORK ARCHITECTURES FOR STYLE CLASSIFICATION

	conv1 (32)	pool1	rnorm1	conv2 (64)	pool2	rnorm2	conv3 (64)	conv4 (32)	conv5 (32)	conv6 (32)	fc1K	fc256	fc14	mAP	Accuracy
Arch 1	√	√	√	√	√	√	√	√			√	√	√	**56.81%**	**59.89%**
Arch 2	√	√	√	√	√	√					√	√	√	52.39%	54.33%
Arch 3	√	√	√	√	√	√	√				√	√	√	53.19%	55.19%
Arch 4	√	√		√	√		√	√	√	√	√	√	√	54.13%	55.77%
Arch 5	√	√	√	√	√	√	√	√				√	√	53.94%	56.00%
Arch 6	√	√	√	√	√	√	√	√	√		√	√	√	53.22%	57.25%
Arch 7	√	√	√	√	√	√	√	√	√	√	√	√	√	47.44%	52.16%

other 7 categories, the SCNN performs better than [10]. We observed that unlike images in the other 7 categories, images in the stilllife category do not share obvious visual similarity. As shown in Figure 10 (left), images in this category involve objects and scenes of various semantic categories that we may encounter in our everyday life, such as animals, fruits, houses, and people. Moreover, we found that images in the stilllife category tend to be associated with certain artistic styles, as shown in Figure 10 (right). This fact makes the problem of assessing the aesthetics of stilllife images more challenging than the other 7 categories using the deep network training approach. Due to the diverse content of images in this category and the limited number of training data (2491), compared to [10], the performance of SCNN in this category is worse in general. An exception, according to Figure 9 (b), is the results produced by "adapt" with the inputs of I_g^w, which slightly outperforms [10].

In our experiments, the adaptation of the SCNN with the input I_l^r takes about 50 epochs, and the adaptation of SCNN with the input I_g^w takes about 10 epochs. The fact that the SCNN adaptation with the input I_g^w requires less epochs is because the random selection of patches l^r results in more training examples than the warping operation g^w.

D. Categorization with Style Attributes

We evaluate the performance of RDCNN in two steps to show the effectiveness of using style attributes in helping image aesthetics prediction. We first evaluate the style classifier, and then we evaluate the aesthetics prediction accuracy achieved by RDCNN.

The style classification performance achieved by SCNN was compared with the performance reported in [10]. The Average Precision (AP) and mean Average Precision (mAP) were used as the evaluation metrics. We trained and evaluated the SCNN on the same collection of training and testing images as presented in [10]. We conducted similar experiments as we have presented in Section III-A to select a best-performed network architecture. We fixed the architecture and compare performance of SCNN using the four input types. As shown in Table IV, the best mAP we achieved is 56.81% which outperforms the accuracy of 53.85% reported in [10]. The best performance is produced by Arch 1, shown in Table IV, using the I_l^r as the input. We visualize the filters learned by the first convolutional layer of SCNN for image style classification in Figure 11.

To demonstrate the effectiveness of style attributes, the RDCNN was compared with DCNN for both $\delta = 0$ and $\delta = 1$. The results, shown in Table III, reveal that RDCNN outperforms DCNN. We qualitatively analyzed the results

Fig. 11. 32 convolutional kernels of the size $11 \times 11 \times 3$ learned by the first convolutional layer of Style-SCNN for style classification.

TABLE V
ACCURACY OF STYLE CLASSIFICATION WITH DIFFERENT INPUTS

	I_l^r	I_g^w	I_g^c	I_g^p
AP	56.93%	44.52%	45.74%	41.78%
mAP	56.81%	47.01%	48.14%	44.07%
Accuracy	59.89%	48.08%	48.85%	46.79%

produced by RDCNN and found that examples correctly classified by $\text{RDCNN}_{\text{style}}$ are mostly associated with obvious stylistic characteristics, such as rule-of-thirds, HDR, black and white, long exposure, complementary colors, vanishing point, and soft focus. Examples are presented in Figure 12. The observation implies that style attributes help image aesthetics prediction in cases when images are associated with obvious styles.

E. Categorization with Semantic Attributes

We evaluate the $\text{RDCNN}_{\text{semantic}}$ approach in the same way with $\text{RDCNN}_{\text{style}}$. In our experiments, we first fine-tuned the ImageNet to the image aesthetics problem by adding an fc256 layer and replacing the label layer. We then trained regularized $\text{RDCNN}_{\text{semantic}}$ as introduced in Section II-C. The categorization results with semantic attributes are shown in Table III, where $\text{RDCNN}_{\text{semantic}}$ improves the accuracy produced by DCNN. By comparing the best aesthetic quality categorization accuracy with and without semantic attributes, we demonstrate the effectiveness of using semantic attributes in determining the aesthetics of images.

We conducted qualitative analysis to analyze the advantage of the RDCNN architecture using semantic attributes. We present the examples in Figure 13, where we show representative test images that have been correctly classified by $\text{RDCNN}_{\text{semantic}}$ but misclassified by DCNN. Our observation is that the classification accuracy improves with images that contain obvious objects.

F. The IAD dataset

We introduce a new large-scale image aesthetics dataset (IAD), containing 1.5 million images, to explore the impact of a larger-scale training dataset to the proposed approach in terms of classification accuracy. Among the images in the IAD dataset, 300K images were derived from the DPChallenge[10],

[10]We crawled all the images on the DPChallenge uploaded upon April 2014.

Fig. 12. Test images correctly classified by $\text{RDCNN}_{\text{style}}$ and misclassified by DCNN. The label on each image indicates the ground truth aesthetic quality of images.

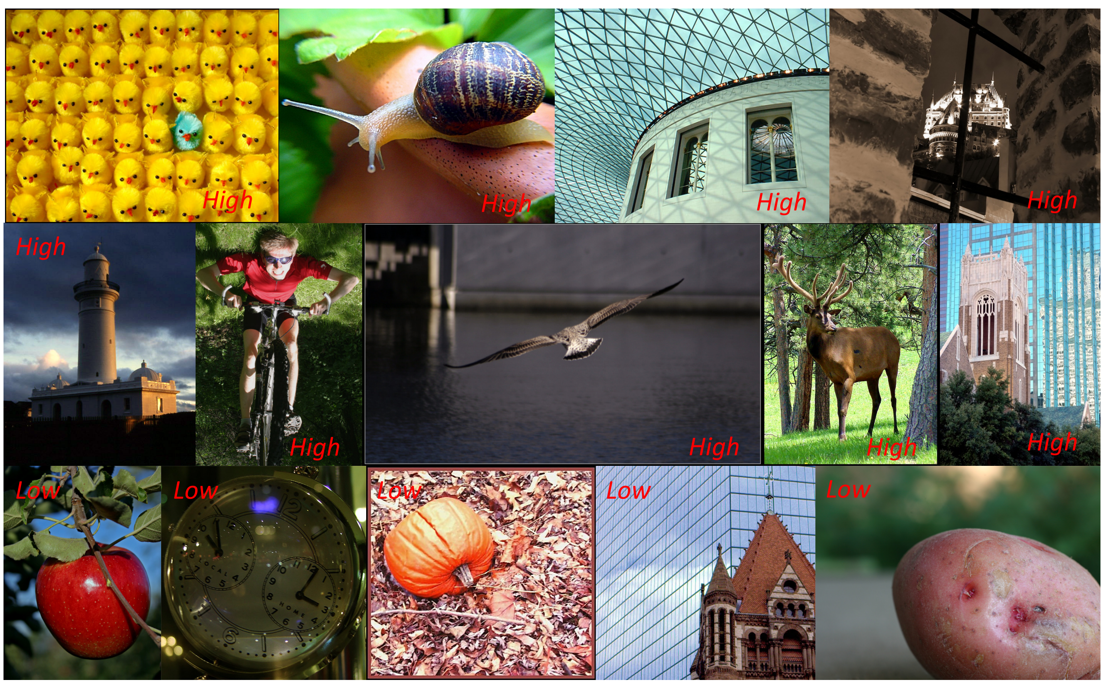


Fig. 13. Test images correctly classified by $\text{RDCNN}_{\text{semantic}}$ and misclassified by DCNN. The label on each image indicates the ground truth aesthetic quality of images.

and 1.2 million were derived from the PHOTO.NET[11]. The score distributions of the two sub-collections are presented in Figure 14.

To generate a training dataset with two categories (low aesthetics and high aesthetics), we divided the images crawled from the PHOTO.NET based on their mean score 4.88 (i.e., images with score higher than 4.88 are labeled as high aesthetics, and images with score lower than 4.88 are labeled as low aesthetics.) For images crawled from DPChallenge, we followed [10] and labeled the images as high aesthetics when the score is larger than 5 and otherwise labeled the images as low aesthetics. We handled the images crawled from the two sources separately because the score scales in the PHOTO.NET $(1-7)$ and the DPChallenge are different $(1-10)$. This results in 747K and 696K training images in the categories of high aesthetics and low aesthetics respectively.

We trained the SCNN and DCNN on the IAD dataset using the same architecture as introduced in [13]. We first trained the SCNN with the input I_l^r, and we evaluated the network on the AVA test set and achieved 73.21% accuracy, about 2% higher than the SCNN trained on AVA dataset. We then trained the SCNN with I_g^w as the input, and the network achieved 73.65% accuracy, 5% percent higher than the SCNN trained on the AVA dataset. We initialized the two columns of DCNN with the SCNN with inputs of the I_l^r and I_g^w, and the DCNN achieved an accuracy of 74.6%, compared to 73.25% using only the AVA training set. Even though the accuracy is not as high as RDCNN using semantic attributes, the results indicate

[11]We included all the images on the PHOTO.NET uploaded upon April 2014 that have been associated with more than 5 aesthetic labels.

that by increasing the size of training data that associated with only aesthetics labels, the prediction accuracy could also be improved.

We attempted an alternative strategy to build the training dataset, that is, using the top 20% rated images as positive samples and the bottom 20% images as negative samples. The accuracy produced by I_g^w and I_l^r are 72.65% and 72.11%, respectively, and the accuracy of DCNN is 72.9%. The results are worse than using the entire IAD dataset for training with the mean value of 4.88 as the boundary for positive and negative training examples. The results may caused by two reasons. First, the AVA test set contains images with ratings in the middle. By cutting off images with medium scores from training, the prediction accuracy on medium-range test images may be affected. Second, utilizing the the top 20% and bottom 20% rated images reduce the number of training data. To fit the new dataset, network architectures have to be carefully adjusted in order to achieve good prediction results, which is non-trivial. We will take it as our future work and discuss the variations of problem formulation in image aesthetics and their corresponding performance.

Beside the large scale, another advantage of the IAD dataset is that a sub-collection of images in the dataset is associated with camera parameters, such as Aperture/FNumber, ISO/ISOSpeedRatings, Shutter/ExposureTime, and Lens/FocalLength. While we did not use these information in this work, we believe such information may facilitate future studies and help users to take aesthetically appealing photos.

G. Implementation Details and Computational Efficiency

All the networks presented in this paper were implemented using ConvNet[12], which supports multi-column inputs for a fully-connected layer. We used the logistic regression cost layer in all network trainings. We initialized the weights and biases learning rate of convolutional and fully-connected layers as 0.001 and 0.002, respectively. Both the weight momentum and the bias momentum were set to 0.9, and the dropout rate was 0.5 on all fully-connected layers. The detailed network architectures are presented in Sections III-A and III-F.

On the AVA dataset, it takes 2 days to train SCNN for a certain input type, and about 3 days for DCNN. Training SCNN for style classification takes roughly a day, and 3-4 days for RDCNN training. With Nvidia Tesla M2070/M2090 GPU, it took about 50 minutes, 80 minutes, and 100 minutes for SCNN, DCNN, and RDCNN, to compute predictions of 2,048 images (each with 50 views) respectively. On the IAD dataset, training SCNN for a specific input type takes about four days, and training DCNN takes about one day using two SCNN as initialization. Classifying 2048 images (each with 50 views) took about 60 minutes, 80 minutes for SCNN and DCNN, respectively, with Nvidia Tesla K40 GPU[13].

IV. Conclusions

This work studied deep neural network training approaches for image aesthetics. In particular, we introduce a double-column deep convolutional neural network approach to assess image aesthetics. Using such novel architecture, we learned aesthetic-related features automatically and unified the feature learning and classifier training. The proposed double-column architecture captures both the global and local views of an image for judging its aesthetic quality. We further developed a network adaptation strategy to apply the trained double-column network training approach for content-based image aesthetics. In addition, image style and semantic attributes are leveraged respectively to boost performance. Experimental results show that our approaches produce significantly higher accuracy than earlier-reported results on the AVA test set, one of the largest existing benchmark with rich aesthetic ratings. Moreover, we introduced a 1.5 million IAD dataset for image aesthetics and improved the aesthetic assessment accuracy on the AVA test set. This result shows that the performance of image aesthetics could be further improved given a larger-scale training dataset.

One limitation of our work is that we have not yet been able to explain what we have learned exactly from the proposed network training approach and why those features help improve the performance. Visualizing trained neuron networks is one of the most active and significant research problems among recent deep learning studies. We would like to treat it as our future work.

[12] https://code.google.com/p/cuda-convnet/

[13] The Nvidia Tesla K40 GPU is faster than Nvidia Tesla M2070/M2090 GPU in testing because the SCNN and DCNN trained on the IAD dataset has much larger capacity than the one trained on the AVA dataset.

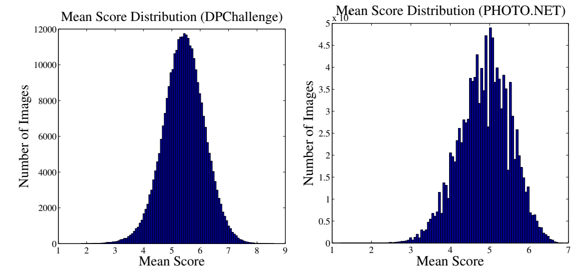


Fig. 14. The Mean Score Distributions of images collected from DPChallenge and PHOTO.NET. Left: The mean score distributions of the 1.2 million PHOTO.NET images; Right: The mean score distributions of the 300K DPChallenge images.

References

[1] R. Datta, D. Joshi, J. Li, and J. Wang, "Studying aesthetics in photographic images using a computational approach," in *European Conference on Computer Vision (ECCV)*, pp. 288–301, 2006.

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

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

[4] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *ACM International Conference on Multimedia (MM)*, pp. 271–280, 2010.

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

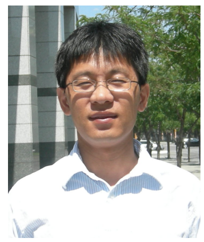
[6] S. Dhar, V. Ordonez, and T. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1657–1664, 2011.

[7] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33–40, 2011.

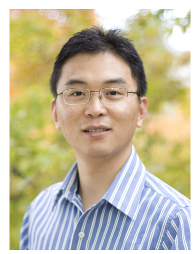
[8] P. O'Donovan, A. Agarwala, and A. Hertzmann, "Color compatibility from large datasets," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 4, pp. 63:1–12, 2011.

[9] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1784–1791, 2011.
[10] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2408–2415, 2012.
[11] L. Marchesotti and F. Perronnin, "Learning beautiful (and ugly) attributes," in *British Machine Vision Conference (BMVC)*, 2013.
[12] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision (IJCV)*, vol. 60, no. 2, pp. 91–110, 2004.
[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1106–1114, 2012.
[14] H.-H. Su, T.-W. Chen, C.-C. Kao, W. Hsu, and S.-Y. Chien, "Scenic photo quality assessment with bag of aesthetics-preserving features," in *ACM International Conference on Multimedia (MM)*, pp. 1213–1216, 2011.
[15] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision (IJCV)*, vol. 42, no. 3, pp. 145–175, 2001.
[16] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3642–3649, 2012.
[17] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," in *The IEEE International Conference on Computer Vision (ICCV)*, 2013.
[18] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage features learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3626–3633, 2013.
[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[20] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
[21] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.
[22] S. Karayev, A. Hertzmann, H. Winnermoller, A. Agarwala, and T. Darrel, "Recognizing image style," in *British Machine Vision Conference (BMVC)*, 2014.
[23] 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," in *Technical report, 2013. arXiv:1310.1531v1*, 2013.
[24] F. Agostinelli, M. Anderson, and H. Lee, "Adaptive multi-column deep neural networks with application to robust image denoising," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1493–1501, 2013.
[25] A. Khosla, A. Das Sarma, and R. Hamid, "What makes an image popular?" in *International World Wide Web Conference (WWW)*, pp. 867–876, 2014.
[26] O. Litzel, in *On Photographic Composition*. New York: Amphoto Books, 1974.
[27] W. Niekamp, "An exploratory investigation into factors affecting visual balance," in *Educational Communication and Technology: A Journal of Theory, Research, and Development*, vol. 29, no. 1, pp. 37–48, 1981.
[28] R. Arnheim, in *Art and visual Perception: A psychology of the creative eye.* Los Angeles. CA: University of California Press., 1974.
[29] D. Joshi, R. Datta, E. Fedorovskaya, Q. T. Luong, J. Z. Wang, J. Li, and J. B. Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, 2011.
[30] J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 22, no. 10, pp. 1345–1359, 2010.
[31] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *International Conference on Machine Learning (ICML)*, pp. 160–167, 2008.
[32] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang, "On shape and the computability of emotions," in *ACM International Conference on Multimedia (MM)*, pp. 229–238, 2012.

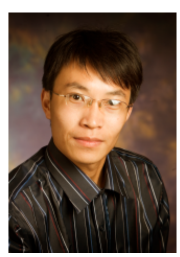
Xin Lu received the B.E. and B.A. degree in Electronic Engineering and English, the M.E. Degree in Signal and Information Processing, all from Tianjin University, China, in 2008 and 2010 respectively. She is currently a Ph.D. candidate at College of Information Sciences and Technology, The Pennsylvania State University, University Park. Her research interests include computer vision and multimedia analysis, deep learning, image processing, Web search and mining, and user-generated content analysis.



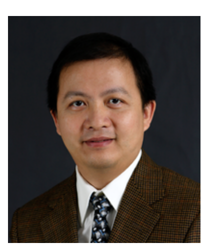
Dr. Zhe Lin received the BEng degree in automatic control from the University of Science and Technology of China in 2002, the MS degree in electrical engineering from the Korea Advanced Institute of Science and Technology in 2004, and the PhD degree in electrical and computer engineering from the University of Maryland, College Park, in 2009. He has been a research intern at Microsoft Live Labs Research. He is currently a senior research scientist at Adobe Research, San Jose, California. His research interests include object detection and recognition, image classification, content-based image and video retrieval, human motion tracking, and activity analysis. He is a member of the IEEE.



Dr. Hailin Jin received his Bachelor's degree in Automation from Tsinghua University, Beijing, China in 1998. He then received his Master of Science and Doctor of Science degrees in Electrical Engineering from Washington University in Saint Louis in 2000 and 2003 respectively. Between fall 2003 and fall 2004, he was a postdoctoral researcher at the Computer Science Department, University of California at Los Angeles. Since October 2004, he has been with Adobe Systems Incorporated where he is currently a Principal Scientist. He received the best student paper award (with J. Andrews and C. Sequin) at the 2012 International CAD conference for work on interactive inverse 3D modeling. He is a member of the IEEE and the IEEE Computer Society.



Dr. Jianchao Yang (S08, M12) received the B.E. degree from the Department of Electronics Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2006, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana, in 2011. He is currently a Research Scientist with the Advanced Technology Laboratory, Adobe Systems Inc., San Jose, CA. His research interests include object recognition, deep learning, sparse coding, compressive sensing, image and video enhancement, denoising, and deblurring.



Dr. James Z. Wang (S96-M00-SM06) is a Professor and the Chair of Faculty Council at the College of Information Sciences and Technology, The Pennsylvania State University. He received a Summa Cum Laude Bachelors degree in Mathematics and Computer Science from University of Minnesota, an M.S. in Mathematics and an M.S. in Computer Science, both from Stanford University, and a Ph.D. degree in Medical Information Sciences from Stanford University. His main research interests are automatic image tagging, aesthetics and emotions, computerized analysis of paintings, and image retrieval.