Yubin Deng, Chen Change Loy, and Xiaoou Tang

Image Aesthetic Assessment

An experimental survey

This article reviews recent computer vision techniques used in the assessment of image aesthetic quality. Image aesthetic assessment aims at computationally distinguishing high-quality from low-quality photos based on photographic rules, typically in the form of binary classification or quality scoring. A variety of approaches has been proposed in the literature to try to solve this challenging problem. In this article, we summarize these approaches based on visual feature types (hand-crafted features and deep features) and evaluation criteria (data set characteristics and evaluation metrics). The main contributions and novelties of the reviewed approaches are highlighted and discussed. In addition, following the emergence of deep-learning techniques, we systematically evaluate recent deep-learning settings that are useful for developing a robust deep model for aesthetic scoring.

Experiments are conducted using simple yet solid baselines that are competitive with the current state of the art. Moreover, we discuss the possibility of manipulating the aesthetics of images through computational approaches. We hope that this article might serve as a comprehensive reference for future research on the study of image aesthetic assessment.

Aesthetic Assessment Through Computer Vision

The aesthetic quality of an image is judged by commonly established photographic rules, which can be affected by numerous factors, including the different uses of lighting [1], contrast [2], and image composition [3] [see Figure 1(a)]. These human judgments, given in an aesthetic evaluation setting, are the result of human aesthetic experience, i.e., the interaction between emotional–valuation, sensory–motor, and meaning–knowledge neural systems, as demonstrated in a systematic neuroscience study by Chatterjee et al. [4]. From the beginning of psychological aesthetics studies by Fechner [5] to modern neuroaesthetics, researchers have argued that there is a certain connection between human aesthetic experience and the sensation caused by visual stimuli, regardless of source, culture, and experience [6], which is supported by activations in specific regions of the visual cortex [7]–[10]. For example, humans' general reward circuitry produces pleasure when they look at beautiful objects [11], and the subsequent aesthetic judgment consists of the appraisal of the valence of such

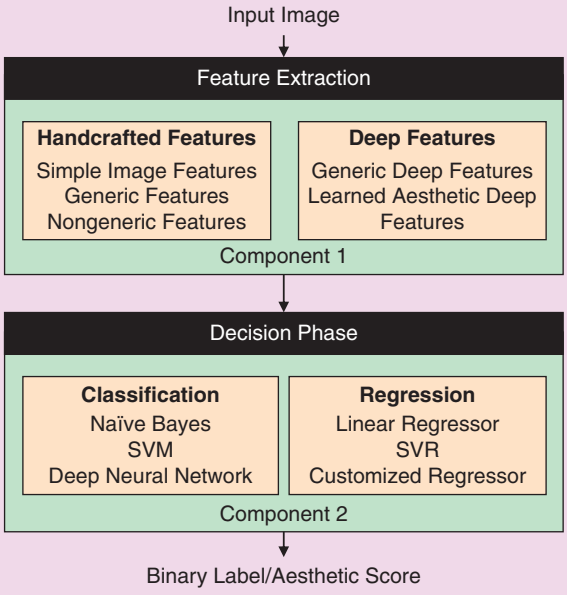
Digital Object Identifier 10.1109/MSP.2017.2696576
Date of publication: 11 July 2017

 1053-5888/17©2017IEEE

perceived objects [8]–[10], [12]. These activations in the visual cortex can be attributed to the processing of various early, intermediate, and late visual features of the stimuli, including orientation, shape, color grouping, and categorization [13]–[16]. Artists intentionally incorporate such features to facilitate desired perceptual and emotional effects in viewers, forming a set of guidelines as they create artworks to induce desired responses in the nervous systems of perceivers [16], [17]. And modern photographers, to make their work appealing to as large an audience as possible, now also resort to certain well-established photographic rules [18], [19] when they capture images.

As the volume of visual data available online grows at an exponential rate, the capability of automatically distinguishing high-quality images from low-quality ones is in increasing demand in real-world image searching and retrieving applications. When a person enters a particular keyword in an image search

(a)



(b)

FIGURE 1. (a) Some high-quality images following well-established photographic rules (top row: color harmony; middle row: single salient object and low depth of field; bottom row: black-and-white portraits with decent lighting contrast). (b) A typical flow of image aesthetic assessment systems. SVM: support vector machine; SVR: support vector regressor.

engine, it is expected that the system will return professional photographs instead of random snapshots. For example, when a user enters the words "mountain scenery," the person will expect to see colorful, pleasing mountain views or well-captured mountain peaks instead of gray or blurry mountain snapshots.

The design of these intelligent systems can potentially be facilitated by insights from neuroscience studies, which show that human aesthetic experience is a kind of information processing that includes five stages: perception, implicit memory integration, explicit classification of content and style, cognitive mastering, and evaluation, which together ultimately produce aesthetic judgment and aesthetic emotion [12], [13]. However, it is nontrivial to computationally model this process. Challenges in the task of judging the quality of an image include 1) computationally modeling the intertwined photographic rules, 2) knowing the aesthetic differences in images from different image genres (e.g., close-shot object, profile, scenery, and night scenes), 3) knowing the type of techniques used in photo capturing (e.g., high-dynamic range, black and white, and depth of field), and 4) obtaining a large amount of human-annotated data for robust testing.

To address these challenges, computer vision researchers typically cast this problem as a classification or regression problem. Early studies started with distinguishing typical snapshots from professional photographs by trying to model the well-established photographic rules using low-level features [20]–[22]. These systems typically involve a training set and a testing set consisting of both high-quality and low-quality images. The system robustness is judged by the model performance on the testing set using a specified metric, such as accuracy. These rule-based approaches are intuitive, as they try to explicitly model the criteria that humans use in evaluating the aesthetic quality of an image. However, more recent studies [23]–[26] have shown that using a data-driven approach is more effective, as the amount of training data available grows from a couple of hundred images to millions. Besides, transfer learning from source tasks with sufficient amounts of data to a target task with relatively fewer training data is also proven feasible, with many successful attempts showing promising results through deep-learning methods [27] with network fine-tuning, where image aesthetics are implicitly learned in a data-driven manner.

As summarized in Figure 1(b), the majority of the aforementioned computer vision approaches for image aesthetic assessment can be categorized based on image representations (e.g., handcrafted features and learned features) and classifiers/regressors training (e.g., support vector machine [SVM] and neural network learning approaches). To the best of our knowledge, no up-to-date survey covers the state-of-the-art methodologies involved in image aesthetic assessment. The last review was published in 2011 by Joshi et al. [28], and no deep learning-based methods were covered. Some reviews on image-quality assessment have been published [29], [30]. In those efforts, image-quality metrics regarding the differences between a noise-tempered sample and the original high-quality image were proposed, including but not limited to mean squared error, structural similarity index (SSIM) [31], and visual information fidelity (VIF) [32]. Nevertheless, their main

focus was on distinguishing noisy images from clean ones in terms of a different quality measure rather than artistic/photographic aesthetics.

In this article, we contribute a thorough overview of the field of image aesthetic assessment. Meanwhile, we also cover the basics of deep-learning methodologies. Specifically, as different data sets exist and evaluation criteria vary in the image aesthetics literature, we do not aim to directly compare the system performance of all of the reviewed works; instead, we point out in the survey their main contributions and novelties in model designs, and give potential insights for future directions in this field of study. In addition, following the recent emergence of deep-learning techniques and the effectiveness of the data-driven approach in learning better image representation, we systematically evaluate different techniques that could facilitate the learning of a robust deep classifier for aesthetic scoring. Our study covers topics such as data preparation, fine-tuning strategies, and multicolumn deep architectures, which we believe to be useful for researchers working in this domain.

In particular, we summarize useful insights on how to alleviate the potential problem of data distribution bias in a binary classification setting and show the effectiveness of rejecting false-positive predictions using our proposed convolutional neural network (CNN) baselines, as revealed by the balanced accuracy metric. We also review the most commonly used publicly available image aesthetic assessment data sets for this problem and draw connections between image aesthetic assessment and image aesthetic manipulation, including image enhancement, computational photography, and automatic image cropping.

Background

The deep neural network

The deep neural network belongs to the family of deep-learning methods that are tasked to learn feature representation in a data-driven approach. While shallow models (e.g., SVM and boosting) showed success in earlier studies concerning relatively smaller amounts of data, they require highly engineered feature designs in solving machine-learning problems. Common architectures in deep neural networks consist of a stack of parameterized individual modules that we call *layers*, such as the convolution layer and the fully connected layer. The architecture design of stacking layers on top of layers is inspired by the hierarchy in the human visual cortex ventral pathway, offering different levels of abstraction for the learned representation in each layer. Information propagation among layers in feed-forward deep neural networks typically follows a sequential pattern. A forward operation $F(\cdot)$ is defined respectively in each layer to propagate the input $\boldsymbol{x}$ it receives and produces an output $\boldsymbol{y}$ to the next layer. For example, the forward operation in a fully connected layer with learnable weights $\mathbf{W}$ can be written as

$$y = F(\boldsymbol{x}) = \mathbf{W}\boldsymbol{x} = \sum w_{ij} \cdot x_i. \tag{1}$$

This is typically followed by a nonlinear function, such as sigmoid

$$z = \frac{1}{1 + \exp(-y)} \tag{2}$$

or the rectified linear unit $z = \max(0, y)$, which acts as the activation function and produces the net activation output z.

To learn the weights $\mathbf{W}$ in a data-driven manner, we need to have the feedback information that reports the current performance of the network. Essentially, we are trying to tune the knobs $\mathbf{W}$ to achieve a learning objective. For example, given an objective t for the input $\boldsymbol{x}$, we want to minimize the squared error between the net output z and t by defining a loss function L:

$$L = \frac{1}{2} \| z - t \|^2. \tag{3}$$

To propagate this feedback information to the weights, we define the backward operation for each layer using gradient backpropagation [33]. We hope to get the direction $\Delta\mathbf{W}$ to update the weights $\mathbf{W}$ to better suit the training objective (i.e., to minimize L): $\mathbf{W} \leftarrow \mathbf{W} - \eta\Delta\mathbf{W}$, where η is the learning rate. In our example, $\Delta\mathbf{W}$ can be easily derived based on the chain rule:

$$\begin{aligned} \Delta\mathbf{W} &= \frac{\partial L}{\partial \mathbf{W}} \\ &= \frac{\partial L}{\partial z} \frac{\partial z}{\partial y} \frac{\partial y}{\partial \mathbf{W}} \\ &= (z - t) \cdot \frac{\exp(-y)}{(\exp(-y) + 1)^2} \cdot \boldsymbol{x}. \end{aligned} \tag{4}$$

In practice, researchers resort to batch stochastic gradient descent or more advanced learning procedures that compute more stable gradients, as averaged from a batch of training examples $\{(\boldsymbol{x}_i, t_i) \mid \boldsymbol{x}_i \in X\}$ to train deeper and deeper neural networks with continually increasing numbers of layers. We refer readers to [27] for an in-depth overview of additional deep-learning methodologies.

Image-quality metrics

Image-quality metrics are defined in an attempt to quantitatively measure the objective quality of an image. This is typically used in image restoration applications (superresolution [34], deblurring [35], and deartifacting [36]), where we have a default high-quality reference image for comparison. However, these quality metrics are not designed to measure the subjective nature of human-perceived aesthetic quality (see examples in Figure 2). Directly applying these objective quality metrics to our domain of image aesthetic assessment may produce misleading results, as can be seen from the measured values in Figure 2(b). Interest in developing more robust metrics has increased in the research community, as a means to assess the more subjective quality of image aesthetics.

A typical pipeline

Most existing image-quality assessment methods take a supervised learning approach. A typical pipeline assumes a set of training data $\{\boldsymbol{x}_i, y_i\}_{i\in[1,N]}$, from which a function $f: g(X) \rightarrow Y$ is learned, where $g(\boldsymbol{x}_i)$ denotes the feature representation of image $\boldsymbol{x}_i$. The label $\boldsymbol{y}_i$ is either $\{0, 1\}$ for binary classification (when f is a classifier) or a continuous score range for regression (when f is a regressor). Following this formulation, a pipeline can be broken into two main components, as shown in Figure 1(b), i.e., a feature extraction component and a decision component.

Feature extraction

The first component of an image aesthetics assessment system aims at extracting robust feature representations describing the aesthetic aspect of an image. Such features are assumed to model the photographic/artistic aspect of images to distinguish images of different qualities. Numerous efforts have been made to design features that are robust enough for the intertwined aesthetic rules. The majority of feature types can be classified into handcrafted features and deep features. Conventional approaches [20], [21], [37]–[49] typically adopt handcrafted features to computationally model the photographic rules (e.g., lighting and contrast), global image layout (the rule of thirds), and typical objects (e.g., human profiles, animals, and plants) in images. In more recent work, generic deep features [50], [51] and learned deep features [23]–[25], [52]–[59] exhibit stronger representation power for this task.

Decision phase

The second component of an image aesthetics assessment system provides the ability to perform classification or regression for the given aesthetic task. The naïve Bayes classifier, SVM, boosting, and deep classifier are typically used for binary classification of high-quality and low-quality images, whereas regressors like support vector regressors (SVRs) are used in ranking or scoring images based on their aesthetic quality.

FIGURE 2. Quality measurements by peak signal-to-noise ratio (PSNR), SSIM [31], and VIF [32] (a higher measurement is better, typically made against a referencing ground-truth high-quality image). Although these are good indicators for measuring the quality of images in image restoration applications, such as the images in (a), they do not reflect human-perceived aesthetic values, as shown by the measurements for the building images in (b).

Data sets

The assessment of image aesthetic quality assumes a standard training set and testing set containing both high-quality and low-quality image examples, as previously mentioned. Judging the ground-truth aesthetic quality of a given image is, however, a subjective task. As such, it is inherently challenging to obtain a large amount of such annotated data. Most of the earlier papers [21], [38], [39] on image aesthetic assessment collect a small amount of private image data. These data sets typically contain from a few hundred to a few thousand images, with binary labels or aesthetic scoring for each image. Yet such data sets where the model performance is evaluated are not publicly available. Much research effort has later been made to contribute publicly available image aesthetic data sets of larger scale for more standardized evaluation of model performance. In the following, we introduce those data sets that are most frequently used in performance benchmarking for image aesthetic assessment.

The Photo.net data set and the DPChallenge data set are introduced in [28] and [60], respectively. These two data sets can be considered the earliest attempts to construct large-scale image databases for image aesthetic assessment. The Photo.net data set contains 20,278 images, with at least ten score ratings per image. The ratings range from zero to seven, with seven assigned to the most aesthetically pleasing photos. Typically, images uploaded to Photo.net are rated as somewhat pleasing, with the peak of the global mean score skewing to the right in the distribution [28]. The more challenging DPChallenge data set contains diverse ratings. The DPChallenge data set contains 16,509 images in total, and was later replaced by the Aesthetic Visual Analysis (AVA) data set, where a significantly larger number of images derived from DPChallenge.com are collected and annotated.

The Chinese University of Hong Kong-PhotoQuality (CUHK-PQ) data set is introduced in [45] and [61]. It contains 17,690 images collected from DPChallenge.com and amateur photographers. All of the images are given binary aesthetic labels and grouped into seven scene categories, i.e., animals, plants, static, architecture, landscape, humans, and night. The standard training and testing set from this data set are random partitions of a 50–50 split or a fivefold cross-validation partition, where the overall ratio of the total number of positive examples and that of the negative examples is around 1:3. Sample images are shown in Figure 3.

The AVA data set [49] contains ~250,000 images in total. These images are obtained from DPChallenge.com and labeled by aesthetic scores. Specifically, each image receives $78 \sim 549$ votes of scores ranging from one to ten. The average score of an image is commonly taken to be its ground-truth label. As such, it contains more challenging examples, as images that lie within the center score range could be aesthetically ambiguous [Figure 4(a)]. For the task of binary aesthetic quality classification, images with an average score higher than a threshold of $5 + \sigma$ are treated as positive examples, and images with an average score lower than $5 - \sigma$ are treated as negative ones. Additionally, the AVA data set contains 14 style attributes and more than 60 category attributes for a subset of images. There are two typical training and testing splits from this data set, i.e., 1) a large-scale standardized partition with ~230,000 training images and ~20,000 testing images using a hard threshold of $\sigma = 0$, and 2) an easier partition modeling that of CUHK-PQ by taking those images whose score ranking is at the top 10% and the bottom 10%, resulting in ~25,000 images for training and ~25,000 images for testing. The ratio of the total number of positive examples to that of the negative examples is around 12:5.

Apart from these two standard benchmarks, more recent research also introduces new data sets that take into consideration the data-balancing issue. The Image Aesthetic Data Set (IAD) introduced in [55] contains 1.5 million images derived from DPChallenge and Photo.net. Similar to AVA, images in the IAD data set are scored by annotators. Positive examples are selected from those images with a mean score larger than a threshold. All IAD images are used for model training, and the model performance is evaluated on AVA in [55]. The ratio of the number of positive examples to that of the negative

FIGURE 3. Some sample images in the CUHK-PQ data set [45]. (a) Distinctive differences can be visually observed between the high-quality (grouped in the green-framed box) and low-quality images (grouped in the red-framed box). (b) The number of images in the CUHK-PQ data set.

examples is around 1.07:1. The Aesthetic and Attributes Database (AADB) [25] also contains a balanced distribution of professional and consumer photos, with a total of 10,000 images. Eleven aesthetic attributes and annotators' IDs are provided. A standard partition with 8,500 images for training, 500 images for validation, and 1,000 images for testing is proposed [25].

The trend toward creating data sets of even larger volume and higher diversity is essential for boosting the research progress in this field of study. To date, the AVA data set serves as a canonical benchmark for performance evaluation of image aesthetic assessment, as it is the first large-scale data set with detailed annotation. Still, the distribution of positive and negative examples in the data set also plays a role in the effectiveness of trained models, as false-positive predictions are as harmful as having a low recall rate in image retrieval and searching applications. In the following, we review major attempts in the literature to build systems for the challenging task of image aesthetic assessment.

The distribution of positive and negative examples in the data set also plays a role in the effectiveness of trained models.

Conventional approaches with handcrafted features

The conventional option for image quality assessment is to hand-design good feature extractors, which requires a considerable amount of engineering skill and domain expertise. Next we review a variety of approaches that exploit hand-engineered features.

Simple image features

Global features are first explored by researchers to model the aesthetic aspect of images. The works by Datta et al. [21] and Ke et al. [37] are among the first to cast aesthetic understanding of images into a binary classification problem. Datta et al. [21] combine low-level and high-level features that are typically used for image retrieval and train an SVM classifier for binary classification of images in terms of aesthetic quality. Ke et al. [37] propose global edge distribution, color distribution, hue count, and low-level contrast and brightness indicators to represent an image; then they train a naïve Bayes classifier based on such features. An even earlier attempt by Tong et al. [20] adopts boosting to combine global low-level simple features (blurriness, contrast, colorfulness, and saliency) to classify professional photographs and ordinary snapshots.

All of these pioneering works present the very first attempts to computationally model the global aesthetic aspect of images using handcrafted features. Even in a recent work, Ayd in

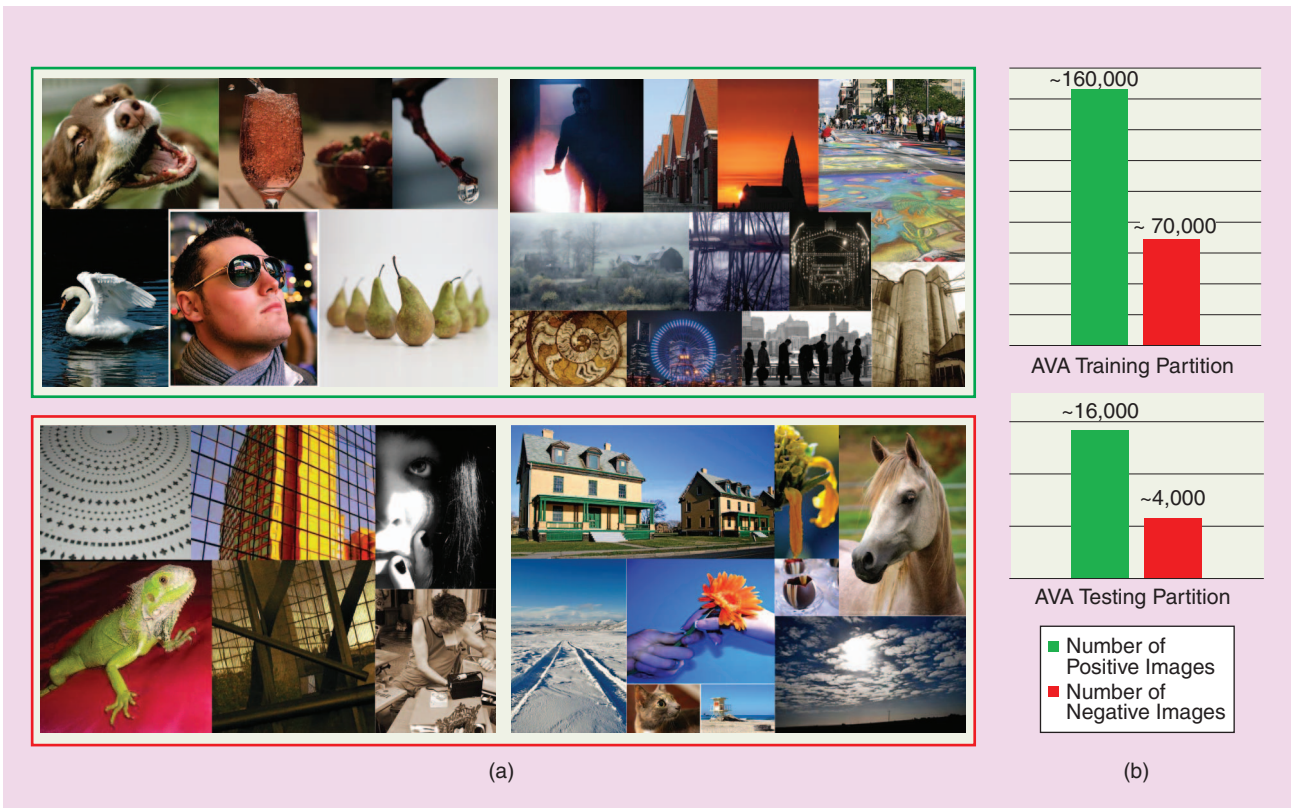


FIGURE 4. Some sample images in the AVA data set [49]. (a) Images in the green-framed box are labeled with a mean score of >5. Images in the red-framed box are labeled with a mean score of <5. The image groups on the right are ambiguous, with a somewhat neutral scoring around five. (b) The number of images in the AVA data set.

et al. [62] construct image aesthetic attributes by sharpness, depth, clarity, tone, and colorfulness. An overall aesthetics rating score is heuristically computed based on these five attributes. Improving upon these global features, later studies adopt global saliency to estimate aesthetic attention distribution. Sun et al. [38] make use of a global saliency map to estimate visual attention distribution to describe an image, and they train a regressor to output the quality score of an image based on the rate-of-focused-attention region in the saliency map. You et al. [39] derive similar attention features based on a global saliency map and incorporate a temporal activity feature for video quality assessment.

Regional image features [40]–[42] later prove to be effective in complementing the global features. Luo et al. [40] extract regional clarity contrast, lighting, simplicity, composition geometry, and color harmony features based on the subject region of an image. Wong et al. [63] compute exposure, sharpness, and texture features on salient regions and global images, as well as features depicting the subject–background relationship of an image. Nishiyama et al. [41] extract bags-of-color patterns from local image regions with a grid-sampling technique. While [40], [41], and [63] adopt the SVM classifier, Lo et al. [42] build a statistical modeling system with coupled spatial relations after extracting color and texture features from images, where a likelihood evaluation is used for aesthetic quality prediction. These methods focus on modeling image aesthetics from local image regions that are potentially most attractive to humans.

The rule of thirds, low depth of field, and opposing colors are the common techniques for composing a good image where the salient object is made outstanding.

Image composition features

Image composition in a photograph typically relates to the presence and position of a salient object. The rule of thirds, low depth of field, and opposing colors are the common techniques for composing a good image where the salient object is made outstanding (see Figure 5). To model such aesthetic aspects, Bhattacharya et al. [43], [64] propose compositional features using relative foreground position and a visual weight ratio to model the relations between foreground objects and the background scene; then an SVR is trained. Wu et al. [65] propose the use of Gabor filter responses to estimate the position of the main object in images, and then extract low-level hue, saturation, value (HSV)-color features from global and central image regions. These features are fed to a soft-SVM classifier with sigmoidal softening to distinguish images of ambiguous quality. Dhar et al. [44] cast high-level features into describable attributes of composition, content, and sky illumination and combine low-level features to train an SVM classifier. Lo et al. [66] propose the combination of layout composition, edge composition features with an HSV color palette, HSV counts, and global features (textures, blur, dark channel, and contrasts). SVM is used as the classifier.

The representative work by Tang et al. [45] gives a comprehensive analysis of the fusion of global features and regional features. Specifically, image composition is estimated by global hue composition and scene composition, and multiple types of regional features extracted from subject areas are proposed, such as dark channel feature, clarity contrast, lighting contrast, composition geometry of the subject region, spatial complexity and human-based features. An SVM classifier is trained on each of the features for comparison, and the final model performance is substantially enhanced by combining all of the proposed features. It is shown that regional features can effectively complement global features in modeling the image aesthetics.

A more recent approach by image composition features is proposed by Zhang et al. [67], where image descriptors that characterize local and global structural aesthetics from multiple visual channels are designed. The spatial structure of the image local regions is modeled using graphlets, and they are connected based on atomic region adjacency. To describe such atomic regions, visual features from multiple visual channels [such as color moment, histogram of oriented gradients (HOG), and saliency histogram] are used. The global spatial layout of the photo is also embedded into graphlets using a Grassmann manifold. The importance of the two kinds of graphlet descriptors is dynamically adjusted, capturing the spatial composition of an image from multiple visual channels. The final aesthetic prediction of an image is generated by a probabilistic model using the postembedding graphlets.

General-purpose features

Yeh et al. [46] make use of scale-invariant feature transform (SIFT) descriptors and propose relative features by matching a query photo to photos in a gallery group. General-purpose imagery features like bag of visual (BOV) words [68] and Fisher vector (FV) [69] are explored in [47]–[49]. Specifically, SIFT and color descriptors are used as the local descriptors upon which a Gaussian mixture model (GMM) is trained. The statistics up to the second order of this GMM distribution are

FIGURE 5. (a) An image composition with low depth of field, a single salient object, and the rule of thirds [49]. (b) An image of low aesthetic quality [45].

then encoded using the BOV words or FV. Spatial pyramid is also adopted, and the per-region encoded FVs are concatenated as the final image representation. These methods ([47]–[49]) represent an attempt to implicitly model photographic rules by encoding them in generic content-based features, which is competitive with or even outperforms simple handcrafted features.

Task-specific features

Task-specific features is a term that refers to features in image aesthetic assessment that are optimized for a specific category of photos, which can be efficient when the use-case or task scenario is fixed or known beforehand. Explicit information (such as human facial characteristics, geometry tag, scene information, or intrinsic character component properties) is exploited based on the different task nature.

Li et al. [70] propose a regression model that targets only consumer photos with faces. Face-related social features (such as facial expression features, facial pose features, and relative facial position features) and perceptual features (facial distribution symmetry, facial composition, and pose consistency) are specifically designed for measuring the quality of images with faces, and it is shown in [70] that for this task they complement conventional handcrafted features (brightness contrast, color correlation, clarity contrast, and background color simplicity). Support vector regression is used to produce aesthetic scores for images.

Lienhard et al. [71] study particular facial features for evaluating the aesthetic quality of headshot images. To design features for face/headshots, the input image is divided into subregions (the eyes, mouth, global face, and entire image regions). Low-level features (sharpness, illumination, contrast, dark channel, and hue and saturation in the HSV color space) are computed from each region. These pixel-level features assume the human way of perceiving a facial image and hence can reasonably model the headshot images. SVM with Gaussian kernel is used as the classifier.

Su et al. [72] propose bag of aesthetics-preserving features for scenic/landscape photographs. Specifically, an image is decomposed into $n \times n$ spatial grids; then low-level features in HSV-color space as well as local binary patterns, HOG, and saliency features are extracted from each patch. The final feature is generated by a predefined patch-wise operation to exploit the landscape composition geometry. AdaBoost is used as the classifier. These features aim at modeling only landscape images and may be limited in their representation power in general image aesthetic assessment.

Yin et al. [73] build a scene-dependent aesthetic model by incorporating the geographic location information with GIST descriptors and spatial layout of saliency features for scene aesthetic classification (such as bridges, mountains, and beaches). SVM is used as the classifier. The geographic location information is used to link a target scene image to relevant photos taken within the same geocontext; then these relevant photos are used as the training partition to the SVM. The authors' proposed model requires input images with geographic tags and is also limited to scenic photos. For scene images without geo-context information, SVM trained with images from the same scene category is used.

Sun et al. [74] design a set of low-level features for aesthetic evaluation of Chinese calligraphy. They target the handwritten Chinese character on a plain white background; hence, conventional color information is not useful in this task. Global shape features, extracted based on standard calligraphic rules, are introduced to represent a character. In particular, the authors consider alignment and stability, distribution of white space, stroke gaps, and a set of component layout features while modeling the aesthetics of handwritten characters. A backpropagation neural network is trained as the regressor to produce an aesthetic score for each given input.

Deep-learning approaches

The powerful feature representation learned from a large amount of data has shown an ever-improving performance in the tasks of recognition, localization, retrieval, and tracking, surpassing the capability of conventional handcrafted features [75]. Since the work by Krizhevsky et al. [75], where CNNs are adopted for image classification, a great degree of interest has arisen in learning robust image representations through deep-learning approaches. Recent works in the literature of image aesthetic assessment using deep-learning approaches to learn image representations can be broken down into two major schemes: 1) adopting generic deep features learned from other tasks and training a new classifier for image aesthetic assessment and 2) learning aesthetic deep features and training a classifier directly from image aesthetics data.

Generic deep features

A straightforward approach to employing deep-learning aims is to adopt generic deep features learned from other tasks and train a new classifier on the aesthetic classification task. Dong et al. [50] propose adopting the generic features from the penultimate layer output of AlexNet [75] with spatial pyramid pooling. Specifically, the $4{,}096\,(\text{fc7}) \times 6\,(\text{SpatialPyramid}) = 24{,}576\text{-dimensional}$ feature is extracted as the generic representation for images; then an SVM classifier is trained for binary aesthetic classification. Lv et al. [51] also adopt the normalized 4,096-dimension fc7 output of AlexNet [75] for feature representation. They propose to learn the relative ordering relationship of images of different aesthetic quality. They use SVM rank [76] to train a ranking model for image pairs of $\{I_{\text{HighQuality}}, I_{\text{LowQuality}}\}$.

Learned aesthetic deep features

Features learned with single-column CNNs

Peng et al. [52] propose to train CNNs of AlexNet-like architecture for eight different abstract tasks (emotion classification, artist classification, artistic style classification, aesthetic classification, fashion style classification, architectural style classification, memorability prediction, and interestingness

prediction). (Figure 6 illustrates a typical single-column CNN.) In particular, the last layer of the CNN for aesthetic classification is modified to output two-dimensional softmax probabilities. This CNN is trained from scratch using aesthetic data, and the penultimate layer (fc7) output is used as the feature representation. To further analyze the effectiveness of the features learned from other tasks, Peng et al. analyze different pretraining and fine-tuning strategies and evaluate the performance of different combinations of the concatenated fc7 features from the eight CNNs.

Wang et al. [53] propose a CNN that is modified from the AlexNet architecture. Specifically, the conv_5 layer of AlexNet is replaced by a group of seven convolutional layers (with respect to different scene categories), which are stacked in a parallel manner with mean pooling before feeding to the fully connected layers, i.e., $\{\text{conv}_5^{1-\text{animal}}, \text{conv}_5^{2-\text{architecture}}, \text{conv}_5^{3-\text{human}}, \text{conv}_5^{4-\text{landscape}}, \text{conv}_5^{5-\text{night}}, \text{conv}_5^{6-\text{plant}}, \text{conv}_5^{7-\text{static}}\}$. The fully connected layers fc6 and fc7 are modified to output 512 feature maps instead of 4,096 for more efficient parameter learning. The 1,000-class softmax output is changed to two-class softmax (fc8) for binary classification. The advantage of this CNN using such a group of seven parallel convolutional layers is to exploit the aesthetic aspects in each of the seven scene categories. During pretraining, a set of images belonging to one of the scene categories is used for each of the $\text{conv}_5^i (i \in \{1,...,7\})$ layers. Then the weights learned through this stage are transferred back to the conv_5^i in the proposed parallel architecture, with the weights from conv_1 to conv_4 reused from AlexNet in the fully connected layer randomly reinitialized. Subsequently, the CNN is further fine-tuned end to end. Upon convergence, the network produces a strong response in the conv_5^i layer feature map when the input image is of category $i \in \{1,...,7\}$. This shows the potential in exploiting image category information when learning the aesthetic presentation.

Tian et al. [54] train a CNN with four convolution layers and two fully connected layers to learn aesthetic features from the data. The output size of the two fully connected layers is set to 16 instead of 4,096 as in AlexNet. The authors propose that such a 16-dimension representation is sufficient to model only the top 10% and bottom 10% of the aesthetic data, which are relatively easy to classify compared to the full data. Based on this efficient feature representation learned from the CNN, the authors propose a query-dependent aesthetic model as the classifier. Specifically, for each query image, a query-dependent training set is retrieved based on predefined rules (visual similarity, image tags association, or a combination of both). Subsequently, an SVM is trained on this retrieved training set. It shows that the features learned from the aesthetic data outperform the generic deep features learned in the ImageNet task.

The deep multipatch aggregation (DMA)-net is proposed in [24], where information from multiple image patches is extracted by a single-column CNN that contains four convolution layers and three fully connected layers, with the last layer outputting a softmax probability. Each randomly sampled image patch is fed into this CNN. To combine multiple feature outputs from the sampled patches of one input image, a statistical aggregation structure is designed to aggregate the features from the orderless sampled image patches by multiple poolings (minimum, maximum, median, and averaging). An alternative aggregation structure is also designed based on sorting. The final feature representation effectively encodes the image based on regional image information.

Features learned from multicolumn CNNs

The Rating Pictorial Aesthetics using Deep Learning (RAPID) model by Lu et al. [23], [55] can be considered to be the first attempt to train CNNs with aesthetic data. They use an AlexNet-like architecture where the last fully connected layer is set to output two-dimensional probability for aesthetic binary classification. Both global image and local image patches are considered in their network input design, and the best model is obtained by stacking a global-column and a local-column CNN to form a double-column CNN, where the feature representation (the penultimate layers' fc7 output) from each column is concatenated before the fc8 layer (classification layer). (Figure 7 shows a typical multicolumn CNN.) Standard stochastic gradient descent is used to train the network with softmax loss. Moreover, the authors further boost the performance of the network by incorporating image style information using a style-column or semantic-column CNN. Then the style-column CNN is used as the third input column, forming a three-column CNN with style/semantic information. Such a multicolumn CNN exploits the data from both the global and local image aspects.

Mai et al. [26] propose stacking five columns of Visual Geometry Group (VGG)-based networks using an adaptive spatial pooling layer. The adaptive spatial pooling layer is

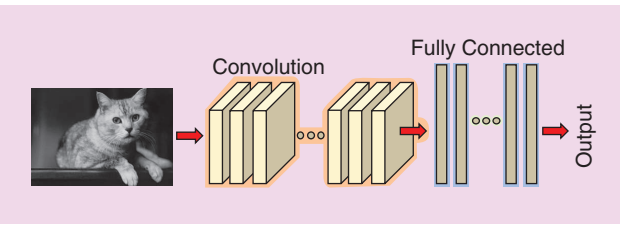


FIGURE 6. The architecture of a typical single-column CNN [49].

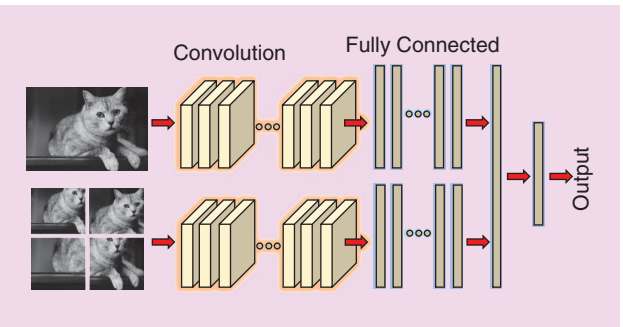


FIGURE 7. A typical multicolumn CNN (a two-column architecture is shown as an example) [49].

designed to allow arbitrary-sized images as input; specifically, it pools a fixed-length output, given different receptive field sizes, after the last convolution layer. By varying the kernel size of the adaptive pooling layer, each subnetwork effectively encodes multiscale image information. Moreover, to potentially exploit the aesthetic aspect of different image categories, a scene categorization CNN outputs a scene category posterior for each input image. Then a final scene-aware aggregation layer processes such aesthetic features (category posterior and multiscale VGG features) and outputs the final classification label. The design of this multicolumn network has the advantage of being able to exploit the multiscale composition of an image in each subcolumn by adaptive pooling, yet the multiscale VGG features may contain redundant or overlapping information, which could potentially lead to network overfitting.

Wang et al. [56] propose a multicolumn CNN model called *brain-inspired deep networks* (*BDN*) that shares similar structures with RAPID. In RAPID, a style attribute prediction CNN is trained to predict 14 style attributes for input images. This attribute CNN is treated as one additional CNN column, which is then added to the parallel input pathways of a global image column and a local patch column. In BDN, 14 different style CNNs are pretrained, and they are parallel cascaded and used as the input to a final CNN for rating distribution prediction, where the aesthetic quality score of an image is subsequently inferred. The BDN model can be considered as an extended version of RAPID that exploits each of the aesthetic attributes using learned CNN features, hence enlarging the parameter space and learning capability of the overall network.

Zhang et al. [57] propose a two-column CNN for learning aesthetic feature representation. The first column (CNN_1) takes image patches as input, and the second column (CNN_2) takes a global image as input. Instead of randomly sampling image patches, given an input image, a weakly supervised learning algorithm is used to project a set of D textual attributes learned from image tags to highly responsive image regions. Such image regions in images are then fed to the input of CNN_1. This CNN_1 contains four convolution layers and one fully connected layer (fc_5) at the bottom. Then a parallel group of D output branches ($\mathrm{fc}_6^i, i \in \{1,2,...,D\}$) modeling each of the D textual attributes are connected on top. The size of the feature maps of each of the fc_6^i is of 128 dimensions. A similar CNN_2 takes a globally warped image as input, producing one more 128-dimension feature vector from fc_6. Hence, the final concatenated feature learned in this manner is $128 \times (D+1)$ dimensional. A probabilistic model containing four layers is trained for aesthetic quality classification.

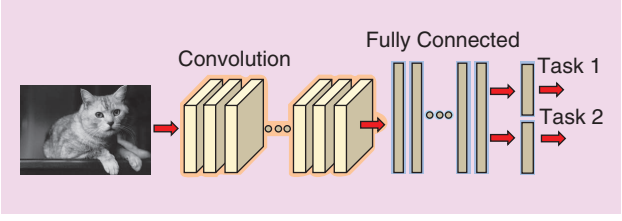


FIGURE 8. A typical multitask CNN consists of a main task (task 1) and multiple auxiliary tasks, only one of which is shown here (task 2) [49].

Kong et al. [25] propose learning aesthetic features assisted by the pair-wise ranking of image pairs as well as the image attribute and content information. Specifically, a Siamese architecture that takes image pairs as input is adopted, where the two base networks of the Siamese architecture adopt the AlexNet configurations (the 1,000-class classification layer fc8 from the AlexNet is removed). In the first stage, the base network is pretrained by fine-tuning from aesthetic data using the Euclidean loss regression layer instead of the softmax classification layer. After that, the Siamese network ranks the loss for every sampled image pair. Upon convergence, the fine-tuned base network is used as a preliminary feature extractor.

In the second stage, an attribute prediction branch is added to the base network to predict image attribute information. Then the base network continues to be fine-tuned in a multitask manner by combining the rating regression Euclidean loss, attribute classification loss, and ranking loss.

In the third stage, yet another content classification branch is added to the base network to predict a predefined set of category labels. Upon convergence, the softmax output of the content category prediction is used as a weighting vector for weighting the scores produced by each feature branch (the aesthetic branch, attribute branch, and content branch).

In the final stage, the base network and all of the added output branches are fine-tuned jointly, with the content classification branch frozen. Effectively, such aesthetic features are learned by considering both the attribute and category content information, and the final network produces image scores for each given image.

Features learned with multitask CNNs

Kao et al. [58] propose three category-specific CNN architectures: one for object, one for scene, and one for texture. The scene CNN takes a warped global image as input. It has five convolution layers and three fully connected layers, with the last fully connected layer producing a two-dimensional softmax classification. The object CNN takes both the warped global image and the detected salient region as input. It is a two-column CNN combining global composition and salient information. The texture CNN takes 16 randomly cropped patches as input. Category information is predicted using a three-class SVM classifier before feeding images to a category-specific CNN. To alleviate the use of the SVM classifier, an alternative architecture with a warped global image as input is trained with a multitask approach, where the main task is aesthetic classification and the auxiliary task is scene category classification. (A typical multitask CNN is illustrated in Figure 8.)

Kao et al. [59] propose learning image aesthetics in a multitask manner. Specifically, AlexNet is used as the base network. Then the 1,000-class fc8 layer is replaced by a two-class

aesthetic prediction layer and a 29-class semantic prediction layer. The loss balance between the aesthetic prediction task and the semantic prediction task is determined empirically. Moreover, another branch containing two fully connected layers for aesthetic prediction is added to the second convolution layer ($conv_2$ of AlexNet). By linking an added gradient flow from the aesthetic task directly to the convolutional layers, one expects to learn better low-level convolutional features. This strategy shares a similar spirit with the deeply supervised net [77].

Evaluation criteria and existing results

Different metrics for performance evaluation of image aesthetic assessment models are used across the literature: classification accuracy [20], [21], [23]–[25], [40], [43], [47], [49], [50], [52]–[59], [63]–[65], [71], [73] reports the proportion of correctly classified results; precision-and-recall (PR) curve [37], [40], [41], [44], [66] considers the degree of relevance of the retrieved items and the retrieval rate of relevant items, which is also widely adopted in image search or retrieval applications; Euclidean distance or residual sum-of-squares error between the ground-truth score and aesthetic ratings [38], [70], [71], [74] and correlation ranking [25], [39], [46] are used for performance evaluation in score regression frameworks; receiver-operating characteristic (ROC) curve [42], [48], [66], [71], [72] and area under the curve [45], [61], [66] concerns the performance of binary classifiers when the discrimination threshold is varied; mean average precision [23], [24], [51], [55] is the average precision (AP) across multiple queries, which is usually used to summarize the PR curve for the given set of samples. These are among the typical metrics for evaluating model effectiveness for image aesthetic assessment (see Table 1 for a summary). Subjective evaluation by conducting human surveys is also seen in [62], where human evaluators are asked to give subjective aesthetic attribute ratings.

We find that it is not feasible to directly compare all methods, as different data sets and evaluation criteria are used across the literature. To this end, we try to summarize, respectively, the released results reported on the two standard data sets, namely the CUHK-PQ (Table 2) and AVA data sets (Table 3), and to present the results on other data sets in Table 4. To date, the AVA data set (standard partition) is considered to be the most challenging by the majority of the reviewed work.

The overall accuracy metric appears to be the most popular metric. It can be written as

$$\text{Overall accuracy} = \frac{TP + TN}{P + N}. \tag{5}$$

This metric alone could be biased and far from ideal, as a naïve predictor that predicts all examples as positive would already reach about $(14k + 0)/(14k + 6k) = 70\%$ classification accuracy. To complement such a metric when evaluating models on imbalanced testing sets, an alternative balanced accuracy metric [78] can be adopted:

$$\text{Balanced accuracy} = \frac{1}{2}\left(\frac{TP}{P}\right) + \frac{1}{2}\left(\frac{TN}{N}\right). \tag{6}$$

Balanced accuracy equally considers the classification performance on different classes [78], [79]. While the overall accuracy in (5) offers an intuitive sense of correctness by reporting the proportion of correctly classified samples, the balanced accuracy in (6) combines the prevalence-independent statistics of sensitivity and specificity. A low balanced accuracy will be observed if a given classifier tends to predict only the dominant class. For the naïve predictor mentioned above, the balanced accuracy would give a proper number

Table 1. An overview of typical evaluation criteria.

Method	Formula	Remarks
Overall accuracy	$\frac{TP+TN}{P+N}$	Accounting for the proportion of correctly classified samples.
Balanced accuracy	$\frac{1}{2}\frac{TP}{P}+\frac{1}{2}\frac{TN}{N}$	Averaging precision and true negative prediction for imbalanced distribution.
PR curve	$p=\frac{TP}{TP+FP}, r=\frac{TP}{TP+FN}$	Measuring the relationship between precision and recall.
Euclidean distance	$\sqrt{\sum_i (Y_i-\hat{Y}_i)^2}$	Measuring the difference between the ground-truth score and aesthetic ratings. Y: ground-truth score, $\hat{Y}$: predicted score.
Correlation ranking	$\frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}}$	Measuring the statistical dependence between the ranking of aesthetic prediction and ground truth. rg_X, rg_Y: rank variables, σ: standard deviation, cov: covariance.
ROC curve	$tpr=\frac{TP}{TP+FN}, fpr=\frac{FP}{FP+TN}$	Measuring model performance change by true positive rate and false positive rate when the binary discrimination threshold is varied.
Mean AP	$\frac{1}{n}\sum_i^n (\text{precision}(i)\times\Delta\text{recall}(i))$	The averaged AP values, based on precision and recall. precision(i) is calculated among the first i predictions, Δrecall(i): change in recall.

TP: true positive, *TN*: true negative, *P*: total positive, *N*: total negative, *FP*: false positive, *FN*: false negative, *tpr*: true positive rate, *fpr*: false positive rate.

Table 2. The methods evaluated on the CUHK-PQ data set.

Method	Data Set	Metric	Result	Training–Testing Remarks
Su et al. (2011) [72]	CUHK-PQ	Overall accuracy	92.06%	1,000 training, 3,000 testing
Marchesotti et al. (2011) [47]	CUHK-PQ	Overall accuracy	89.90%	50–50 split
Zhang et al. (2014) [67]	CUHK-PQ	Overall accuracy	90.31%	50–50 split, 12,000 subset
Dong et al. (2015) [50]	CUHK-PQ	Overall accuracy	91.93%	50–50 split
Tian et al. (2015) [54]	CUHK-PQ	Overall accuracy	91.94%	50–50 split
Zhang et al. (2016) [57]	CUHK-PQ	Overall accuracy	88.79%	50–50 split, 12,000 subset
Wang et al. (2016) [53]	CUHK-PQ	Overall accuracy	92.59%	4:1:1 partition
Lo et al. (2012) [66]	CUHK-PQ	Area under ROC curve	0.93	50–50 split
Tang et al. (2013) [45]	CUHK-PQ	Area under ROC curve	0.9209	50–50 split
Lv et al. (2016) [51]	CUHK-PQ	Mean AP	0.879	50–50 split

Table 3. The methods evaluated on the AVA data set.

Method	Data Set	Metric	Result	Training–Testing Remarks
Marchesotti et al. (2013) [48]	AVA	ROC curve	*tpr*: 0.7, *fpr*: 0.4	Standard partition
AVA handcrafted features (2012) [49]	AVA	Overall accuracy	68.00%	Standard partition
Spatial pyramid pooling (SPP) (2015) [24]	AVA	Overall accuracy	72.85%	Standard partition
RAPID (full method) (2014) [23]	AVA	Overall accuracy	74.46%	Standard partition
Peng et al. (2016) [52]	AVA	Overall accuracy	74.50%	Standard partition
Kao et al. (2016) [58]	AVA	Overall accuracy	74.51%	Standard partition
RAPID (improved version) (2015) [55]	AVA	Overall accuracy	75.42%	Standard partition
DMA-net (2015) [24]	AVA	Overall accuracy	75.41%	Standard partition
Kao et al. (2016) [59]	AVA	Overall accuracy	76.15%	Standard partition
Wang et al. (2016) [53]	AVA	Overall accuracy	76.94%	Standard partition
Kong et al. (2016) [25]	AVA	Overall accuracy	77.33%	Standard partition
BDN (2016) [56]	AVA	Overall accuracy	78.08%	Standard partition
Zhang et al. (2014) [67]	AVA	Overall accuracy	83.24%	10% subset, 12.5k*2
Dong et al. (2015) [50]	AVA	Overall accuracy	83.52%	10% subset, 19k*2
Tian et al. (2016) [54]	AVA	Overall accuracy	80.38%	10% subset, 20k*2
Wang et al. (2016) [53]	AVA	Overall accuracy	84.88%	10% subset, 25k*2
Lv et al. (2016) [51]	AVA	Mean AP	0.611	10% subset, 20k*2

indication of $0.5 \times (14k/14k) + 0.5 \times (0k/6k) = 50\%$ performance on AVA.

In this regard, in the following sections where we discuss our findings on a proposed strong baseline, we report both overall classification accuracy and balanced accuracy to get a more reasonable measure of baseline performance.

Experiments on deep-learning settings

It is evident from Table 3 that deep learning-based approaches dominate the performance of image aesthetic assessment. The effectiveness of learned deep features in this task has motivated us to take a step back to consider how a CNN works to understand the aesthetic quality of an image. It is worth noting that training a robust deep aesthetic scoring model is nontrivial, and often we found that the devil is in the details. To this end, we design a set of systematic experiments based on a baseline one-column CNN and a two-column CNN, and evaluate different settings from minibatch formation to complex multicolumn architecture. The results are reported on the widely used AVA data set.

We observe that by carefully training the CNN architecture, the two-column CNN baseline reaches comparable or

Table 4. The methods evaluated on other data sets.

Method	Data Set	Metric	Result
Tong et al. (2004) [20]	29,540-image private set	Overall accuracy	95.10%
Datta et al. (2006) [21]	3,581-image private set	Overall accuracy	75%
Sun et al. (2009) [38]	600-image private set	Euclidean distance	3.5135
Wong et al. (2009) [63]	3,161-image private set	Overall accuracy	79%
Bhattacharya (2010, 2011) [43], [64]	~650-image private set	Overall accuracy	86%
Li et al. (2010) [70]	500-image private set	Residual sum-of-squares error	2.38
Wu et al. (2010) [65]	10,800-image private set from Flickr	Overall accuracy	~83%
Dhar et al. (2011) [44]	16,000-image private set from DPChallenge	PR curve	–
Nishiyama et al. (2011) [41]	12,000-image private set from DPChallenge	Overall accuracy	77.60%
Lo et al. (2012) [42]	4,000-image private set	ROC curve	*tpr*: 0.6, *fpr*: 0.3
Yeh et al. (2012) [46]	309-image private set	Kendalls Tau-b measure	0.2812
Aydin et al. (2015) [62]	955-image subset from DPChallenge.com	Human survey	–
Yin et al. (2012) [73]	13,000-image private set from Flickr	Overall accuracy	81%
Lienhard et al. (2015) [71]	Human Face Scores 250-image data set	Overall accuracy	86.50%
Sun et al. (2015) [74]	1,000-image Chinese handwriting	Euclidean distance	–
Kong et al. (2016) [25]	AADB data set	Spearman ranking	0.6782
Zhang et al. (2016) [57]	PNE	Overall accuracy	86.22%

even better performance than state-of-the-art methods, and the one-column CNN baseline acquires the strong capability to suppress false-positive predictions while having competitive classification accuracy. We hope the experimental results will facilitate the design of future deep-learning models for image aesthetic assessment.

Formulation and the base CNN structure

The supervised CNN learning process involves a set of training data $\{\boldsymbol{x}_i, y_i\}_{i\in[1,N]}$ from which a nonlinear mapping function $f: X \rightarrow Y$ is learned through backpropagation [33]. Here, $\boldsymbol{x}_i$ is the input to the CNN and $y_i \in \mathbb{T}$ is its corresponding ground-truth label. For the task of binary classification, $y_i \in \{0,1\}$ is the aesthetic label corresponding to image $\boldsymbol{x}_i$. The convolutional operations in such a CNN can be expressed as

$$F_k(X) = \max\left(\mathbf{w}_k * F_{k-1}(X) + \mathbf{b}_k, 0\right), k \in \{1,2,...,D\}, \quad (7)$$

where $F_0(X) = X$ is the network input and D is the depth of the convolutional layers. The operator $*$ denotes the convolution operation. The operations in the D' fully connected layers can be formulated in a similar manner. To learn the $(D + D')$ network weights $\mathbf{W}$ using the standard backpropagation with stochastic gradient descent, we adopt the cross-entropy classification loss, which is formulated as

$$\begin{aligned} L(\mathbf{W}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{t}\{t \log\ p(\hat{y}_i = t \mid \boldsymbol{x}_i; \mathbf{W}) \\ + (1-t)\log\left(1 - p(\hat{y}_i = t \mid \boldsymbol{x}_i; \mathbf{W})\right) + \phi(\mathbf{W})\} \end{aligned} \quad (8)$$

$$p(\hat{y}_i = t \mid \boldsymbol{x}_i; \mathbf{w}_t) = \frac{\exp\left(\mathbf{w}_t^T \boldsymbol{x}_i\right)}{\sum_{t' \in \mathbb{T}} \exp\left(\mathbf{w}_t^T \boldsymbol{x}_i\right)}, \quad (9)$$

where $t \in \mathbb{T} = \{0,1\}$ is the ground truth. This formulation is in accordance with prior successful model frameworks, such as AlexNet [75] and VGG-16 [80], which are also adopted as the base network in some of our reviewed approaches.

The original last fully connected layer of these two networks is for the 1,000-class ImageNet object recognition challenge. For aesthetic quality classification, a two-class aesthetic classification layer to produce a softmax predictor is needed [see Figure 9(a)]. Following typical CNN approaches, the input size is fixed to $224 \times 224 \times 3$, which is cropped from globally warped $256 \times 256 \times 3$ images. Standard data augmentation, such as mirroring, is performed. All of the baselines are implemented based on the Caffe package [81]. For clarity of presentation in the following sections, we name all of our fine-tuned baselines Deep Aesthetic Net (DAN), with the corresponding suffix.

Training from scratch versus fine-tuning

Fine-tuning from a trained CNN has been proven in [36] and [83] to be an effective initialization approach. The RAPID base network [23] uses global image patches and trains a network structure from scratch that is similar to AlexNet. For a fair comparison of similar-depth networks, we first select AlexNet pretrained with the ILSVRC-2012 training set (1.2 million images) and fine-tune it with the AVA training partition. As

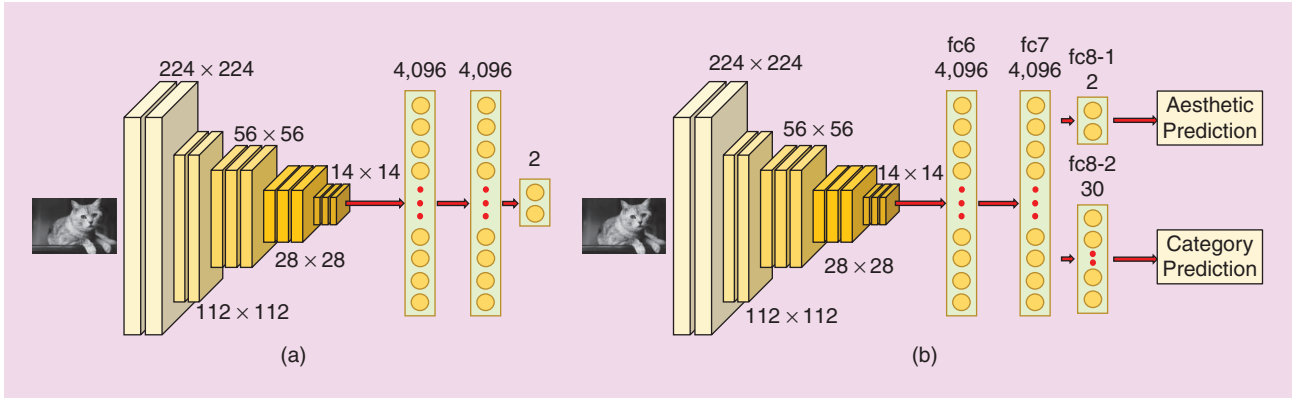


FIGURE 9. (a) The structure of the chosen base network for our systematic study on aesthetic quality classification. (b) The structure of the one-column CNN baseline with multitask learning [49].

Table 5. Training from scratch versus fine-tuning.

Method	Balanced Accuracy	Overall Accuracy
RAPID (global) [23]	–	67.8
DAN-1 (fine-tuned from AlexNet)	68.0	71.3
DAN-1 (fine-tuned from VGG-16)	72.8	74.1

Using a one-column CNN baseline (DAN-1) fine-tuned on AlexNet and VGG-16, both of which are pretrained on the ImageNet data set. The authors in [23] have not released detailed classification results.

Table 6. The effects of minibatch formation.

Minibatch Formation	Balanced Accuracy	Overall Accuracy
DAN-1 (randomly sampled)	70.39	77.65
DAN-1 (balanced formation)	72.82	74.06

Using a one-column CNN baseline (DAN-1) with VGG-16 as the base network.

shown in Table 5, fine-tuning from the vanilla AlexNet yields better performance than simply training the RAPID base network from scratch. Moreover, the DAN model fine-tuned from VGG-16 [see Figure 9(a)] yields the best performance in both balanced accuracy and overall accuracy. It is worth pointing out that other more recent and deeper models, such as ResNet [84], Inception-ResNet [85], and PolyNet [86], could serve as pretrained models. Nevertheless, owing to the typically small size of aesthetic data sets, precautions need be taken during the fine-tuning process. Plausible methods include freezing some earlier layers to prevent overfitting [83].

Minibatch formation

Minibatch formation directly affects the gradient direction toward which stochastic gradient descent brings down the training loss in the learning process. We consider two types of minibatch formation and reveal the impact of this difference on image aesthetic assessment.

Random sampling

By randomly selecting examples for minibatches [87], [88], we select from a distribution of the training partition. Since the number of positive examples in the AVA training partition is almost twice that of the negative examples [Figure 4(b)], models trained with such minibatches may bias toward predicting positives.

Balanced formation

Another approach is to enforce a balanced number of positives and negatives in each of the minibatches, i.e., for each iteration of backpropagation, the gradient is computed from a balanced number of positive examples and negative examples.

Table 6 compares the performance of these two strategies. We observe that although the model fine-tuned with randomly sampled minibatches reaches a higher overall accuracy, its performance is inferior to the one fine-tuned with balanced minibatches, as evaluated using balanced accuracy. To keep track of both true-positive prediction rates and true-negative prediction rates, balanced accuracy is adopted to measure the model robustness on the data imbalance issue. Network fine-tuning in the rest of the experiments is performed with balanced minibatches, unless otherwise specified.

Triplet pretraining and multitask learning

Apart from directly training using the given training data pairs $\{\boldsymbol{x}_i, y_i\}_{i\in[1,N]}$, one could utilize richer information inherent in the data or auxiliary sources to enhance the learning performance. We discuss two popular approaches next.

Pretraining using triplets

The triplet loss is inspired by Dimensionality Reduction by Learning an Invariant Mapping [89] and large margin nearest neighbor [90]. It is widely used in many recent vision studies [79], [91]–[93] and aims to bring data of the same class closer while moving data of different classes further away. This loss is particularly suitable to our task; i.e., the absolute aesthetic score of an image is arguably subjective, but the general relationship that beautiful images are close to each other while the opposite images should be apart is obvious.

Table 7. Triplets pretraining and multitask learning.

Methods	Balanced Accuracy	Overall Accuracy
DAN-1	72.82	74.06
DAN-1 (triplet pretrained)	73.29	75.32
DAN-1 (multitask–aesthetic and category)	73.39	75.36
DAN-1 (triplet pretrained + multitask)	**73.59**	74.42

Using a one-column CNN baseline (DAN-1) with VGG-16 as the base network. Balanced minibatch formation is used.

To enforce such a relationship in an aesthetic embedding, one needs to generate minibatches of triplets for deep feature learning, i.e., an anchor $\boldsymbol{x}$, a positive instance $\boldsymbol{x}_{+ve}$ of the same class, and a negative instance $\boldsymbol{x}_{-ve}$ of a different class. Furthermore, we found it useful to constrain each image triplet to be selected from the same image category. In addition, we observed better performance by introducing triplet loss in the pretraining stage and continuing with conventional supervised learning on the triplet-pretrained model. Table 7 shows that the DAN model pretrained with triplets gives better performance.

We further visualize some categories in the learned aesthetic embedding space in Figure 10. It is interesting to observe that the embedding learned with triplet loss demonstrates much better aesthetic grouping in comparison to that without the use of triplet loss.

Multitask learning with image category prediction

Can aesthetic prediction be facilitated provided that a model understand to which category the image belongs? Following the work in [94], where auxiliary information is used to regularize the learning of the main task, we investigate the potential benefits of using image categories as an auxiliary label in training the aesthetic quality classifier.

Specifically, given an image labeled with main task label y, where $y = 0$ for low-quality images and $y = 1$ for high-quality ones, we provide an auxiliary label $c \in C$ denoting one of the image categories, such as animals, landscape, portraits, and so forth. In total, we include 30 image categories. To learn a classifier for the auxiliary class, a new fully connected layer is attached to the fc7 of the vanilla VGG-16 structure to predict a softmax probability for each category class. The modified one-column CNN baseline architecture is shown in Figure 9(b). The loss function in (8) is now changed to

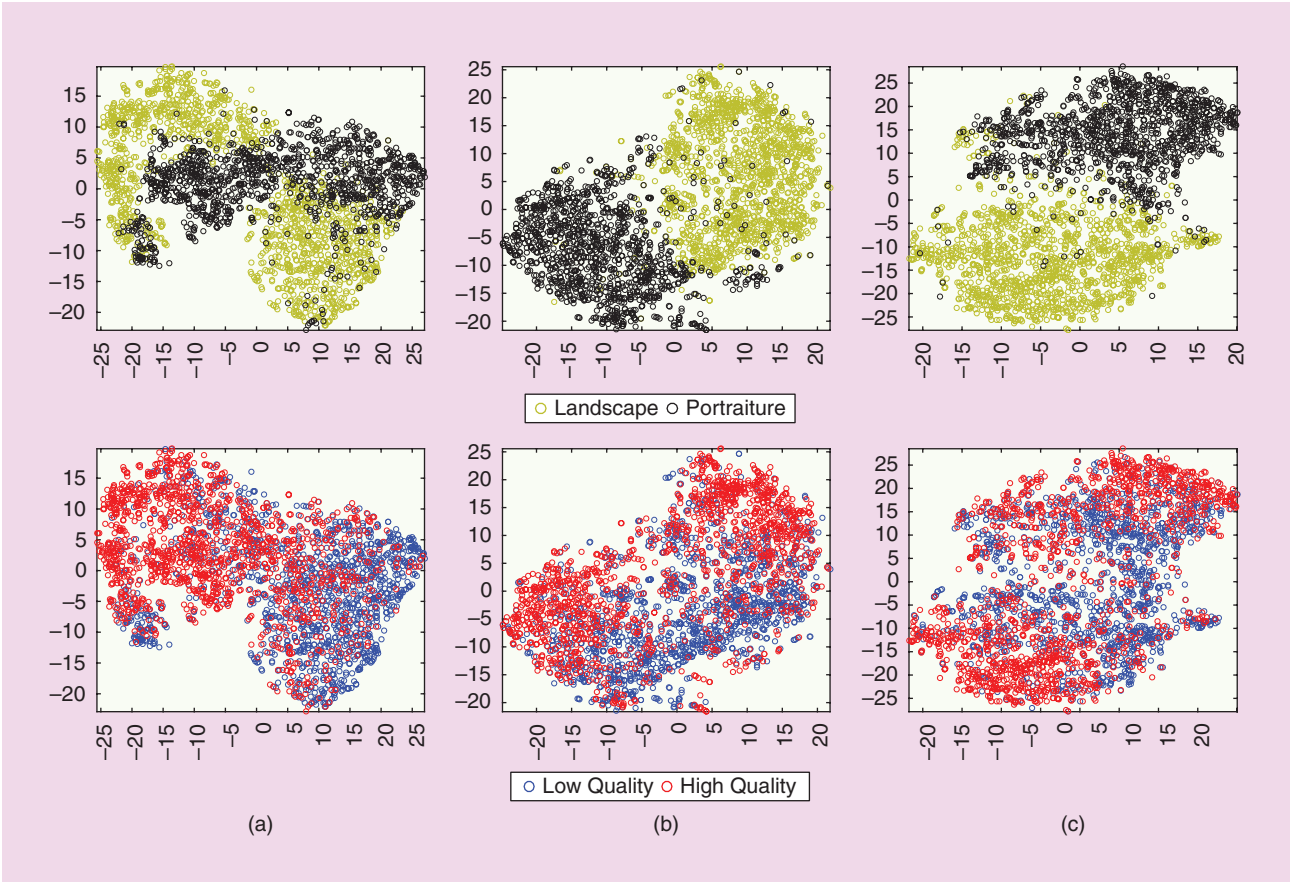


FIGURE 10. Aesthetic embeddings of AVA images (testing partition) learned by triplet loss, visualized using t-SNE [84]: (a) ordinary supervised learning without triplet pretraining and multitask learning, (b) triplet pretrained, and (c) combined triplet pretraining and multitask learning. t-SNE: t-distributed stochastic neighbor embedding.

$$L_{\text{multitask}} = L(\mathbf{W}) + L_{\text{aux}}(\mathbf{W}_c), \tag{10}$$

$$L_{\text{aux}}(\mathbf{W}_c) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C}\{t_c \log p(\hat{y}_c^{\text{aux}} = t_c \mid \boldsymbol{x}_i; \mathbf{W}_c) + (1 - t_c)\log p(\hat{y}_c^{\text{aux}} = t_c \mid \boldsymbol{x}_i; \mathbf{W}_c)) + \phi(\mathbf{W}_c)\}, \tag{11}$$

where $t_c \in \{0, 1\}$ is the binary label corresponding to each auxiliary class $c \in C$ and $\hat{y}_c^{\text{aux}}$ is the auxiliary prediction from the network. Solving the above loss function, the DAN model performance from this multitask learning strategy is observed to have surpassed the previous one (Table 7). It is worth noting that the category annotation of the AVA-training partition is not complete, with about 25% of the images not having categories labeled. For those training instances without categories labeled, the auxiliary loss $L_{\text{aux}}(\mathbf{W}_c)$ due to missing labels is ignored.

The absolute aesthetic score of an image is arguably subjective, but the general relationship that beautiful images are close to each other while the opposite images should be apart is obvious.

Triplet pretraining + multitask learning

Combining triplet pretraining and multitask learning, the final one-column CNN baseline reaches a balanced accuracy of 73.59% on the challenging task of aesthetic classification. The results for different fine-tuning strategies is summarized in Table 7.

Discussion

Note that it is nontrivial to boost the overall accuracy at the same time as we try not to overfit the baseline to a certain data distribution. Still, compared with other released results in Table 8, with careful training, a one-column CNN baseline yields a strong capability of rejecting false positives while attaining a reasonable overall classification accuracy. We show some qualitative classification results as follows.

Figures 11 and 12 show the qualitative results of aesthetic classification by the one-column CNN baseline, using DAN-1 (triplet pretrained + multitask). Note that these examples are correctly classified neither by BDN [56] nor by DMA-net [24]. False-positive test examples (Figure 13) by the DAN-1 baseline still show a somewhat high-quality image trend, with high color contrast or depth of field, while false-negative testing examples (Figure 14) mostly reflect low image tones. Both quantitative and qualitative results suggest the importance of minibatch formation and fine-tuning strategies.

Multicolumn deep architecture

State-of-the-art approaches [23], [24], [55], [56] for image aesthetic classification typically adopt multicolumn CNNs (Figure 7) to enhance the learning capacity of the model. In particular, these approaches benefit from learning multiscale image information (e.g., global image versus local patches) or utilizing image semantic information (e.g., image styles). To incorporate insights from previous successful approaches, we prepared another two-column CNN baseline (DAN-2) (see Figure 15) with a focus on the more apparent approach of using local image patches as a parallel input column. Both [23] and [24] utilize CNNs trained with local image patches as alternative columns in their multibranch network, with performance evaluated using overall accuracy. For fair comparison, we prepared local image patches of size $224 \times 224 \times 3$ following [23] and [24], and we fine-tuned one DAN-1 model from the vanilla VGG-16 (ImageNet) with such local patches. Another branch is the original DAN-1 model, fine-tuned with globally warped input by triplet pretraining and multitask learning (see the section "Triplet Pretraining and Multitask Learning"). We performed separate experiments where minibatches of these local image patches were taken from either random sampling or the balanced formation.

As shown in Table 8, the DAN-1 model fine-tuned with local image patches performs less well under the metric of balanced accuracy compared to the original DAN-1 model fine-tuned with globally warped input in both random minibatch learning and balanced minibatch learning. We conjecture that local patches contain

Table 8. A comparison of aesthetic quality classification between our proposed baselines and previous state-of-the-art methods on the canonical AVA testing partition.

Previous Work	Balanced Accuracy	Overall Accuracy
AVA handcrafted features (2012) [49]	–	68.00
SPP (2015) [24]	–	72.85
RAPID (full method) (2014) [23]	–	74.46
Peng et al. (2016) [52]	–	74.50
Kao et al. (2016) [58]	–	74.51
RAPID (improved version) (2015) [55]	61.77	75.42
DMA-net (2015) [24]	62.80	75.41
Kao et al. (2016) [59]	–	76.15
Wang et al. (2016) [53]	–	76.94
Kong et al. (2016) [25]	–	77.33
Mai et al. (2016) [26]	–	77.40
BDN (2016) [56]	67.99	78.08
Proposed Baseline Using Random Minibatches		
DAN-1 (VGG-16, AVA global warped input)	70.39	77.65
DAN-1 (VGG-16, AVA local patches)	68.70	77.60
Two-column DAN-2	69.45	**78.72**
Proposed Baseline Using Balanced Minibatches		
DAN-1 (VGG-16, AVA global warped input)	**73.59**	74.42
DAN-1 (VGG-16, AVA local patches)	71.40	75.8
Two-column DAN-2	73.51	75.96

The authors of [23]–[26], [49], [52], [53], [55], [58], and [59] have not released detailed results.

FIGURE 11. Some positive examples (high-quality images) that are wrongly classified by BDN and DMA-net but correctly classified by the DAN-1 baseline [49].

FIGURE 12. Some negative examples (low-quality images) that are wrongly classified by BDN and DMA-net but correctly classified by the DAN-1 baseline [49].

FIGURE 13. Some examples with a negative ground truth that are wrongly classified by the DAN-1 baseline. High color contrast or depth of field is observed in these testing cases [49].

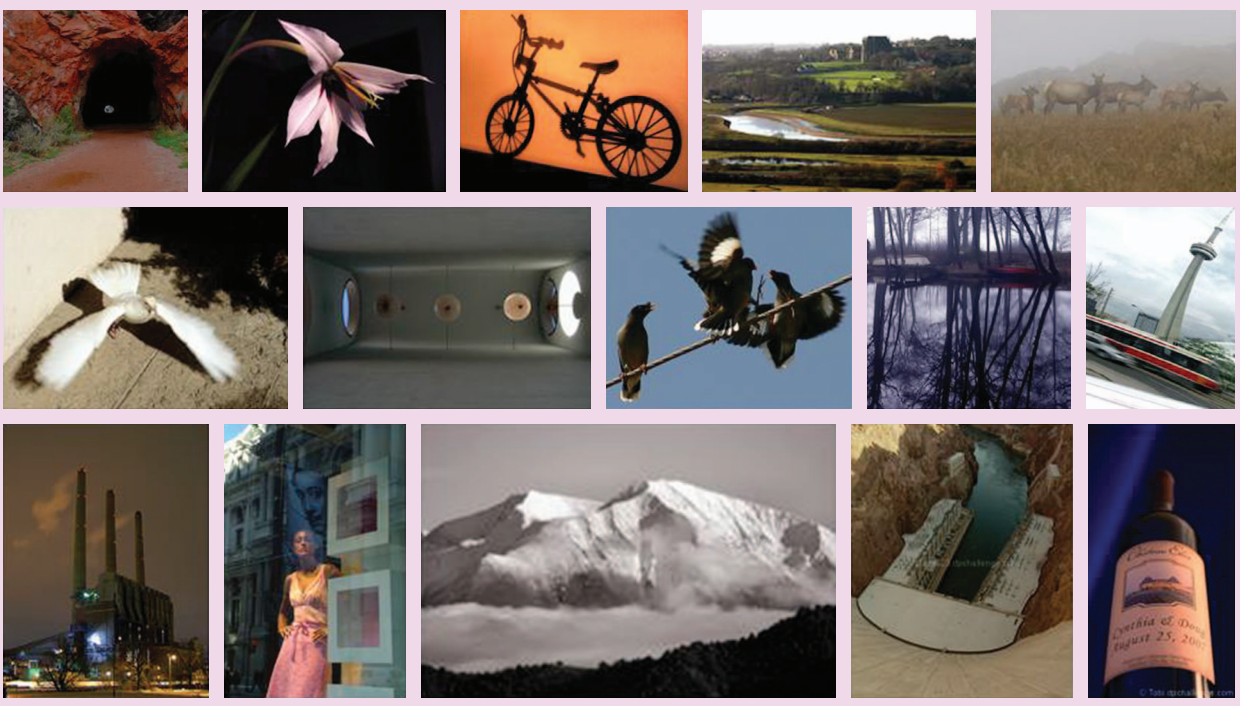


FIGURE 14. Some examples with a positive ground truth that are wrongly classified by the DAN-1 baseline. Most of these images are of low image tones [49].

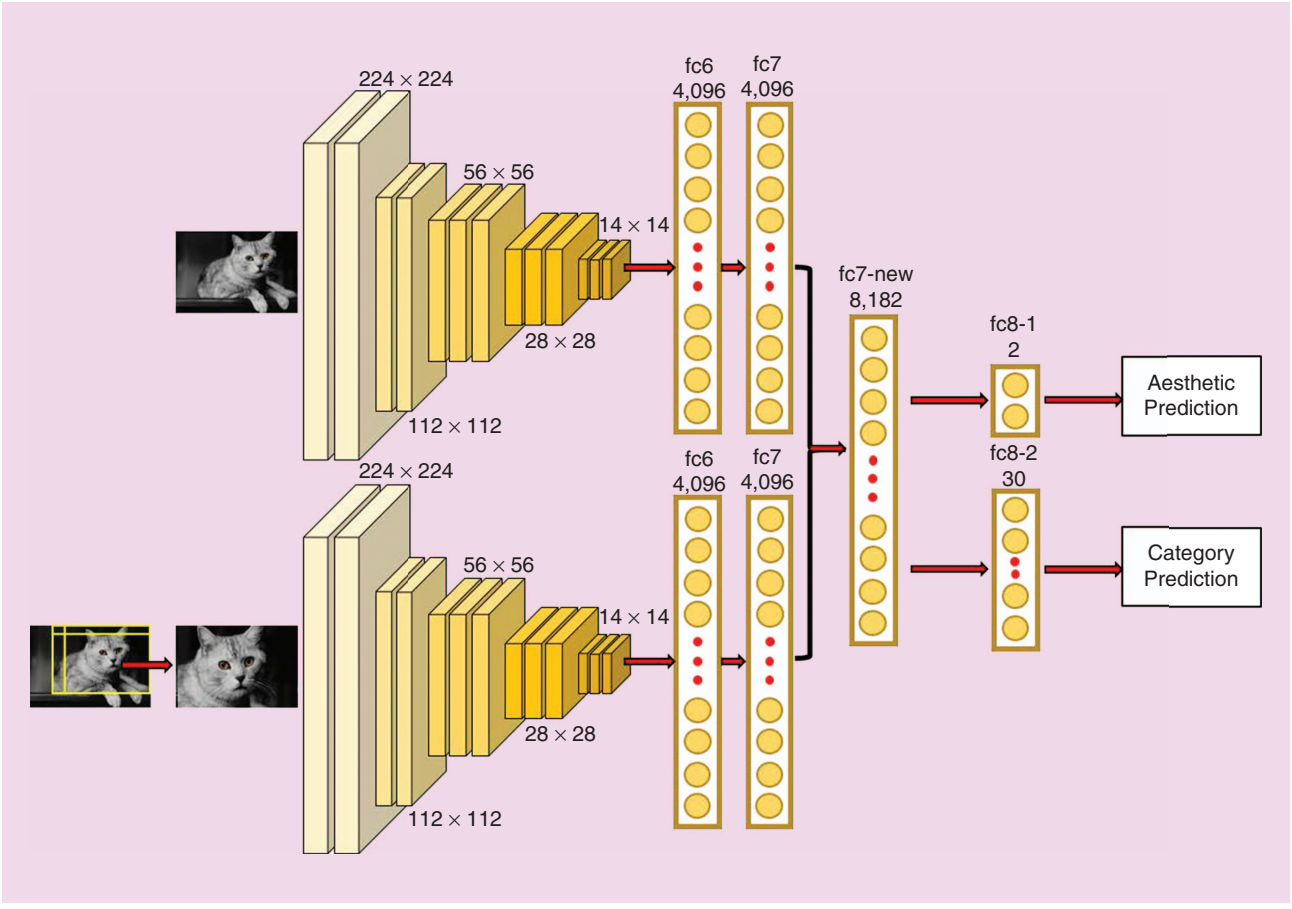


FIGURE 15. The structure of the two-column CNN baseline with multitask learning [49].

no global and compositional information as compared to globally warped input. Nevertheless, such a drop in accuracy is not observed under the overall accuracy metric.

We next evaluated the two-column CNN baseline DAN-2 using the DAN-1 model fine-tuned with local image patches and the one fine-tuned with globally warped input. We have two variants here, depending on whether we employ random or balanced minibatches. We observed that DAN-2 trained with random minibatches attains the highest overall accuracy on the AVA standard testing partition compared to the previous state-of-the-art methods (see Table 8). (Some other works [50], [54], [95]–[97] on AVA data sets use only a small subset of images for evaluation, which is not directly comparable to the canonical state of the art on the AVA standard partition; see Table 3).

Can aesthetic prediction be facilitated provided that a model understand to which category the image belongs?

Interestingly, we observed the balanced accuracy of the two variants of DAN-2 degrades when compared to the respective DAN-1 trained on globally warped input. This observation raises the question of whether local patches necessarily benefit the performance of image aesthetic assessment. We analyzed the cropped local patches more carefully and found that these patches were inherently ambiguous. Thus, the model trained with such inputs could easily become biased toward predicting local patch input to be of high quality, which also explains the performance differences in the two complementary evaluation metrics.

Model depth and layer-wise effectiveness

Determining the aesthetics of images from different categories takes varying photographic rules. We understand that it is not easy to determine some image genres' aesthetic quality in general. It would be interesting to perform a layer-by-layer analysis and track to what degree a deep model has learned image aesthetics in its hierarchical structure. We conducted this experiment using the one-column CNN baseline DAN-1 (triplet pretrained + multitask). We used layer features generated by this baseline model and trained an SVM classifier to perform aesthetic classification on the AVA testing images and then evaluated the performance of different layer features across different image categories.

Features extracted from the convolutional layers of the model were aggregated into a convolutional Fisher representation, as done in [98]. Specifically, to extract features from the dth convolutional layer, note that the output feature maps of

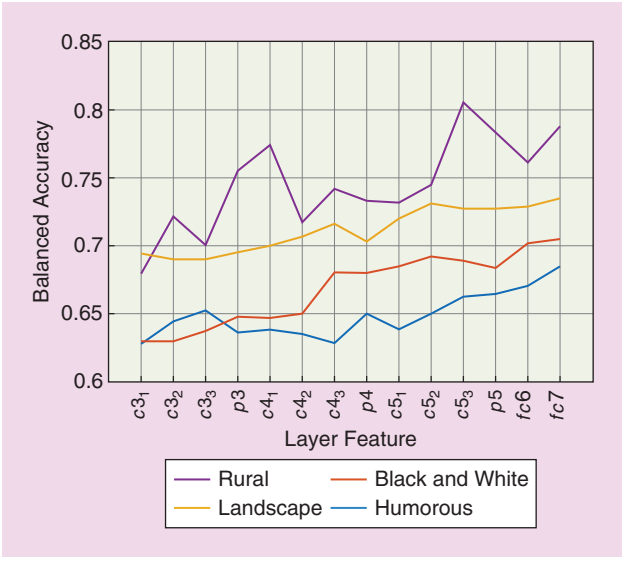


FIGURE 16. A layer-by-layer analysis showing the difficulties of understanding aesthetics across different categories. From the learned feature hierarchy and the classification results, we observe that image aesthetics in the landscape and rural categories can be judged reasonably by the proposed baselines, yet the more ambiguous humorous and black-and-white images are inherently difficult for the model to handle (see also Figure 17).

this dth layer are of size $w \times h \times K$, where $w \times h$ is the size of each of the K output maps. Denote M^k as the kth output map. Specifically, a point $M^k_{i,j}$ in output map M^k is computed from a local patch region L of the input image I using the forward propagation. By aligning all such points into a vector $\boldsymbol{v}_L = [M^1_{i,j}, M^2_{i,j}, ..., M^k_{i,j}, ..., M^K_{i,j}]$, we obtained the feature representation of the local patch region L. A dictionary codebook was created using GMM from all of the $\{\boldsymbol{v}_L\}_{L \in I_{\text{train}}}$, and an FV representation is subsequently computed using this codebook to describe an input image. The obtained convolutional Fisher representation is used for training SVM classifiers.

We compared features from layer conv3_1 to fc7 of the DAN-1 baseline and reported selected results that we find interesting in Figure 16. We obtained the following results:

1) *Model depth is important*: More abstract aesthetic representation can be learned in deeper layers. The performance of aesthetic assessment can generally be benefited from model depth. This observation aligns with that in general object recognition tasks.
2) *Different categories demand different model depths*: The aesthetic classification accuracy on images belonging to the black and white category are generally lower than the accuracy on images in the landscape category across all of the layer features. Sample classification results are shown in confusion matrix ordering (see Figure 17). High-quality black-and-white images show subtle details that should be considered when assessing their aesthetic level, whereas high-quality landscape images differentiate from those low-quality ones in a more apparent way. Similar observations are found, e.g., in the humorous and rural categories. The observation explains why it could be inherently difficult for the baseline model to judge whether images from some specific categories are aesthetically pleasing or not, revealing yet another challenge in the assessment of image aesthetics.

DAN-2 trained with random minibatches attains the highest overall accuracy on the AVA standard testing partition compared to the previous state-of-the-art methods.

From generic aesthetics to user-specific taste

Individual users may hold different opinions on the aesthetic quality of any single image. One may consider that all of the images in Figure 13 are of high quality to some extent, even though the average scores by the data set annotators say otherwise. Coping with individual aesthetic bias is a challenging problem. We may follow the idea behind transfer learning [83] and directly model the aesthetic preference of individual users by transferring the learned aesthetic features to fitting personal taste. In particular, we consider that the DAN-1 baseline network has already captured a sense of generic aesthetics in the aforementioned learning process; so to adapt to personal aesthetic preferences, one can include additional data sources for positive training samples that are user specific, such as the user's personal photographic album or the collection of photos that the user "liked" on social media. As such, our proposed baseline can be further fine-tuned with personal-taste data for individual users and become a personalized aesthetic classifier.

Image aesthetic manipulation

A task closely related to image aesthetic assessment is image aesthetics manipulation, the aim of which is to improve the aesthetic quality of an image. A full review of the techniques of image aesthetics manipulation in the literature is beyond the scope of this article. Still, we make an attempt to connect image aesthetic assessment to a broader topic surrounding image aesthetics by focusing on one of the major aesthetic enhancement operations, i.e., automatic image cropping.

Aesthetics-based image cropping

Image cropping improves the aesthetic composition of an image by removing undesired regions, increasing its aesthetic value. A majority of cropping schemes in the literature can be divided into three main approaches. Attention/saliency-based approaches [99]–[101] typically extract the primary subject region in the scene of interest according to attention scores or saliency maps as the image crops. Aesthetics-based approaches [102]–[104] assess the attractiveness of some proposed candidate crop windows with low-level image features and rules of photographic composition. However, simple handcrafted features are not robust for modeling the huge aesthetic space. The state-of-the-art method is the change-based approach proposed by Yan et al. [105], [106], which aims to

FIGURE 17. A layer-by-layer analysis of classification results using the best layer features on (a) black-and-white category images and (b) landscape category images [49].

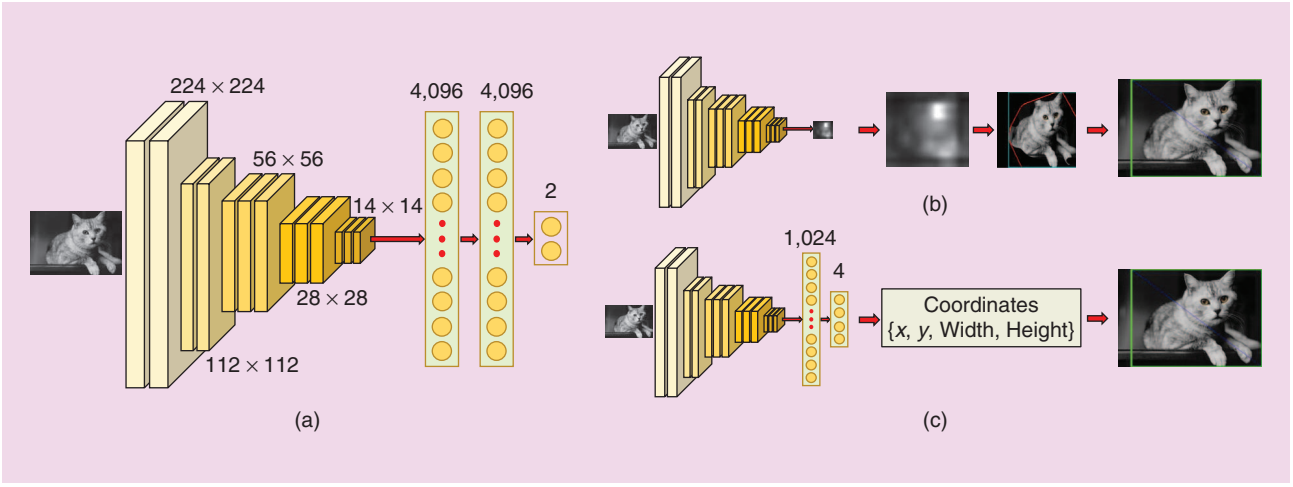


FIGURE 18. (a) The originally proposed one-column CNN baseline. (b) A tweaked CNN made by removing all of the fully connected layers. (c) A modified CNN incorporating a crop-regression layer to learn cropping coordinates [49].

account for what is removed and changed by cropping itself and trying to incorporate the influence of the starting composition of the initial image in the ending composition of the cropped image. This approach produces reasonable crop windows, but the time cost of producing an image crop is prohibitively expensive because of the time spent in evaluating large numbers of crop candidates.

Automatic thumbnail generation is also closely related to automatic image cropping. Huang et al. [107] target visual representativeness and foreground recognizability when cropping and resizing an image to generate its thumbnail. Chen et al. [108] aim at extracting the most visually important region as the image crop. Nevertheless, the aesthetics aspects of cropping are not taken into prime consideration in these approaches.

In the next section, we show that high-quality image crops can already be produced from the last convolutional layer of the aesthetic classification CNN. Optionally, this convolutional response can be utilized as the input to a cropping regression layer for learning more precise cropping windows from additional crop data.

High-quality image crops can already be produced from the last convolutional layer of the aesthetic classification CNN.

Plausible formulations based on deep models

Fine-tuning a CNN model for the task of aesthetic quality classification (see the "Experiments on Deep-Learning Settings" section) can be considered as a learning process in which the fine-tuned model tries to understand the metric of image aesthetics. We hypothesize that the same metric is applicable to the task of automatic image cropping. We discuss two possible variants as follows.

DAN-1 (original) without cropping data

Without utilizing additional image cropping data, a CNN such as the one-column CNN baseline DAN-1 can be tweaked to produce image crops with minor modifications, removing the fully connected layers. That leaves us with a neural network that is fully convolutional where the input can be of arbitrary size, as shown in Figure 18(b). The output of the last convolutional layer of the modified model is $14 \times 14 \times 512$ dimensional, where the 512 feature maps contain the responses/activations corresponding to the input. To generate the final image crop, we take an average of the 512 feature maps and resize it to the input image size. After that, a binary mask is generated by suppressing the feature map values below a threshold. The output crop window is produced by taking a rectangle convex hull from the largest connected region of this binary mask.

DAN-1 (regression) with cropping data

Alternatively, to include additional image cropping data $\{\boldsymbol{x}_i^{\text{crop}}, Y_i^{\text{crop}}\}_{i\in[1,N]}$, where $Y_i^{\text{crop}} = [x, y, \text{width}, \text{height}]$, we follow insights in [111] and add a window regression layer to learn a mapping from the convolutional response [see Figure 18(c)]. As such, we can predict a more precise cropping window by learning this extended regressor from such crop data by a Euclidean loss function:

$$L(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} \| \hat{Y}_i^{\text{crop}} - Y_i^{\text{crop}} \|^2, \tag{12}$$

where $\hat{Y}_i^{\text{crop}}$ is the predicted crop window for input image $\boldsymbol{x}_i^{\text{crop}}$.

To learn the regression parameters for this additional layer, the image cropping data set by Yan et al. [105] is used for further fine-tuning. Images in the data set are labeled with ground-truth crops by professional photographers. Following the evaluation criteria in [105], a fivefold cross-validation approach is adopted for evaluating the model performance on

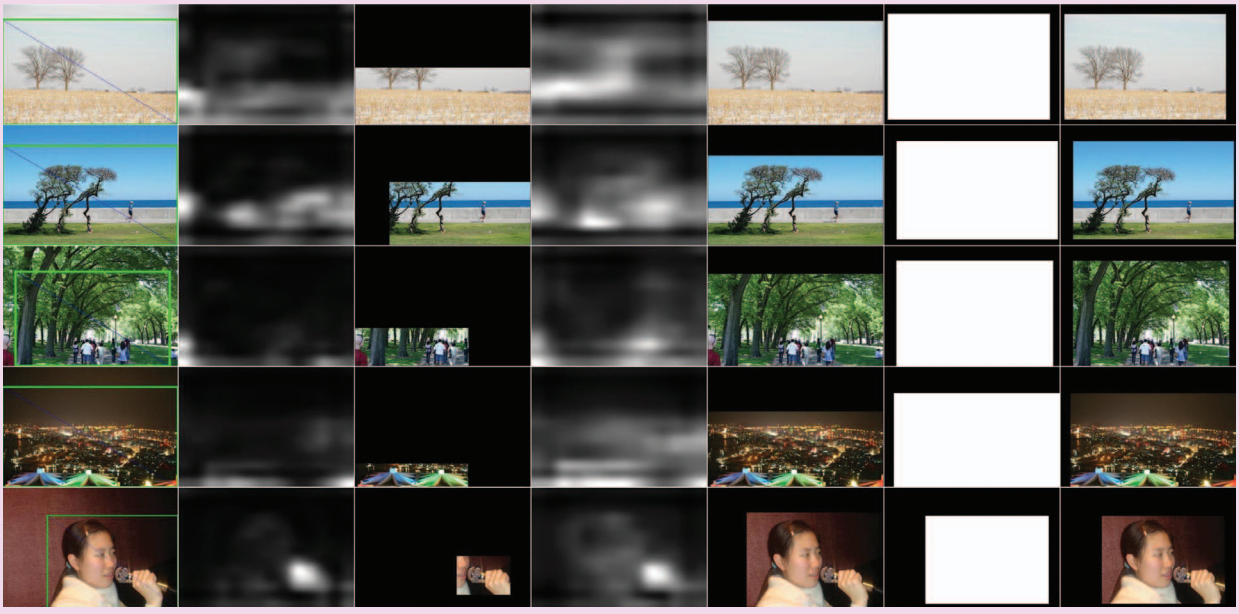


FIGURE 19. The layer response differences of the last convolutional layer. The images in each row correspond to (a) the input image with ground-truth crop, (b) the feature response of the vanilla VGG, (c) the image crops obtained via the feature responses of the vanilla VGG, (d) the feature response of the DAN-1 (original) model, (e) the image crops obtained via the DAN-1 (original) model, (f) the four-coordinates window estimated by the DAN-1 (regression) network, and (g) the cropped image generated by the DAN-1 (regression) [107].

all images in the data set. Note that there are only a few hundred images in each training fold; hence, a direct fine-tuning by simply warping the few hundred images of input to $224 \times 224 \times 3$ could be vulnerable to overfitting. To this end, we fix the weights in the convolutional layers of the DAN-1 (regression) network and learn only the weights for the crop window regression layers. Also, a systematic augmentation approach is adopted as follows. First, the input images are randomly jittered for a few pixels ($\times 5$), and mirroring is performed ($\times 2$). Second, we warp the images to have their longer side equal to 224, hence keeping their aspect ratio. We further downscale the images using a scale of $C \in \{50\%, 60\%, 80\%, 90\%\}$ ($\times 4$). The downscaled images are then padded back to 224 × 224 from $\{\text{top} - \text{left}, \text{top} - \text{right}, \text{bottom} - \text{left}, \text{bottom} - \text{right}\}$ ($\times 4$). Finally, we have direct input warping regardless of the aspect ratio ($\times 1$). In this manner, one training instance is augmented to $5 \times 2 \times (4 \times 4 + 1) = 170$ input instances. We fine-tune this modified CNN baseline with a learning rate of $10e^{-3}$, and the fine-tuning process converges at around the second epoch.

Aesthetics-based image cropping

As shown in Figure 19, we observe that the convolutional response of the vanilla VGG-16 (ImageNet) for object recognition typically finds a precise focus of the salient object in view, while the one-column CNN baseline, i.e., the DAN-1 (original) for aesthetic quality classification, outputs an aesthetically oriented salient region where both the object in view and its object composition are revealed. Compared to the cropping performance using the vanilla VGG-16, image crops from our DAN-1 (original) baseline already have the capability of removing unwanted regions while preserving the aesthetically salient part in view (see Figure 19). The modified CNN, i.e., the DAN-1 (regression), further incorporates aesthetic composition information in its crop window regression layer, which serves to refine the crop coordinates for more precise crop generation.

Following the evaluation settings in [105] and [106], we use the average overlap ratio and average boundary displacement error to quantify the performance of automatic image cropping. A higher overlap and a lower displacement between the generated crop and the corresponding ground truth indicate a more precise crop predictor. As shown in Table 9, directly using the DAN-1 (original) baseline responses to construct image crops already gains competitive cropping performance, while fine-tuning the DAN-1 (regression) with cropping data further boosts the performance and even surpasses the previous state-of-the-art method [105] on this data set, especially in terms of the boundary displacement error. Last but not least, it is worth noting that the CNN-based cropping approach takes merely ~0.2 s for generating an output image crop on a graphics processing unit and ~2 s on a central processing unit (compared to ~11 s on CPU in [105]).

Table 9. The performance on automatic image cropping.

Previous Work	*Photographer 1	Photographer 2	Photographer 3
Park et al. [111]	0.6034 (0.1062)	0.5823 (0.1128)	0.6085 (0.1102)
Yan et al. [108]	0.7487 (0.0667)	0.7288 (0.0720)	0.7322 (0.0719)
Wang et al. [112]	0.7823 (0.0623)	0.7697 (0.0617)	0.7725 (0.0701)
Yan et al. [107]	0.7974 (0.0528)	**0.7857** (0.0567)	0.7723 (0.0594)
Proposed Baselines			
Vanilla VGG-16 (ImageNet)	0.6971 (0.0580)	0.6841 (0.0618)	0.6715 (0.0613)
DAN-1 (original) (AVA training partition)	0.7637 (0.0437)	0.7437 (0.0493)	0.7360 (0.0495)
DAN-1 (regression) (cropping data fine-tuned)	**0.8059 (0.0310)**	0.7750 **(0.0375)**	**0.7725 (0.0377)**

*There are separate ground-truth annotations by three different photographers in the cropping data set of [107].
The first number is the average overlap ratio (higher is better). The second number (shown in parentheses) is the average boundary displacement error (lower is better).
Bold values signify the best performance by the corresponding methods.

Conclusion and potential directions

Models with competitive performance on image aesthetic assessment have been seen in the literature, yet the state of research in this field is far from saturated. Challenging issues include the ground-truth ambiguity due to neutral image aesthetics and how to effectively learn category-specific image aesthetics from the limited amount of auxiliary data information. Image aesthetic assessment can also benefit from an even larger volume of data, with richer annotations, where every single image is labeled by more users with diverse backgrounds. A large and more diverse data set will facilitate the learning of future models and potentially allow more meaningful statistics to be captured.

In this work, we systematically review major attempts on image aesthetic assessment in the literature and further propose an alternative baseline to investigate the challenging problem of understanding image aesthetics. We also discuss an extension of image aesthetic assessment to the application of automatic image cropping by adapting the learned aesthetic-classification CNN for the task of aesthetics-based image cropping. We hope that this survey can serve as a comprehensive reference source and inspire future research in understanding image aesthetics and fostering many potential applications.

Authors

Yubin Deng (dy015@ie.cuhk.edu.hk) received his B.Eng. degree (first-class honors) in information engineering from the Chinese University of Hong Kong in 2015. He is currently working toward his Ph.D. degree in the Department of Information Engineering, Chinese University of Hong Kong, with a Hong Kong Ph.D. Fellowship. His research interests include computer vision, pattern recognition, and machine learning. He was a Hong Kong Jockey Club Scholar in 2013–2014. He received the Professor Charles K. Kao Student Creativity Awards champion award in 2015.

Chen Change Loy (ccloy@ie.cuhk.edu.hk) received his B.Eng degree (first-class honors) from the University of Science, Malaysia, in 2005 and his Ph.D. degree in computer science from Queen Mary University of London, United Kingdom, in 2010. He is currently a research assistant professor in the Department of Information Engineering, Chinese University of Hong Kong. Previously, he was a postdoctoral researcher at Queen Mary University of London and Vision Semantics Ltd. His research interests include computer vision and pattern recognition, with a focus on facial analysis, deep learning, and visual surveillance. He serves as an associate editor of *IET Computer Vision Journal* and is a guest editor of *Computer Vision and Image Understanding.* He is a Member of the IEEE.

Xiaoou Tang (xtang@ie.cuhk.edu.hk) received his B.S. degree from the University of Science and Technology of China, Hefei, in 1990, his M.S. degree from the University of Rochester, New York, in 1991, and his Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, in 1996. He is a professor in and the chair of the Department of Information Engineering, Chinese University of Hong Kong. He worked as the group manager of the Visual Computing Group at Microsoft Research Asia from 2005 to 2008. His research interests include computer vision, pattern recognition, and video processing. He received the Best Paper Award at the IEEE Conference on Computer Vision and Pattern Recognition 2009. He was a program chair of the IEEE International Conference on Computer Vision 2009, and he is an editor-in-chief of *International Journal of Computer Vision* and an associate editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence*. He is a Fellow of the IEEE.

References

[1] M. Freeman, *The Complete Guide to Light and Lighting in Digital Photography* (A Lark Photography Book). New York: Sterling Publishing Company, 2007.

[2] J. Itten, *Design and Form: The Basic Course at the Bauhaus and Later.* New York: Wiley, 1975.

[3] B. London and J. Upton, *Photography.* London: Pearson, 2005.

[4] A. Chatterjee and O. Vartanian, "Neuroscience of aesthetics," *Ann. New York Acad. Sci*., vol. 1369, no. 1, pp. 172–194, 2016.

[5] G. T. Fechner, *Vorschule der Aesthetik*, vol. 1. Wiesbaden, Germany: Breitkopf & Härtel, 1876.

[6] S. Zeki, "Clive Bell's 'significant form' and the neurobiology of aesthetics," *Frontiers Human Neurosci*., vol. 7, p. 730, Nov. 2013.

[7] T. Ishizu and S. Zeki, "The brain's specialized systems for aesthetic and perceptual judgment," *Euro. J. Neurosci.*, vol. 37, no. 9, pp. 1413–1420, 2013.

[8] S. Brown, X. Gao, L. Tisdelle, S. B. Eickhoff, and M. Liotti, "Naturalizing aesthetics: Brain areas for aesthetic appraisal across sensory modalities," *Neuroimage*, vol. 58, no. 1, pp. 250–258, 2011.

[9] L. F. Barrett, B. Mesquita, K. N. Ochsner, and J. J. Gross, "The experience of emotion," *Annu. Rev. Psychol.*, vol. 58, pp. 373–403, Jan. 2007.

[10] L. F. Barrett and T. D. Wager, "The structure of emotion: Evidence from neuroimaging studies," *Current Directions Psychol. Sci.*, vol. 15, no. 2, pp. 79–83, 2006.

[11] S. Kühn and J. Gallinat, "The neural correlates of subjective pleasantness," *Neuroimage*, vol. 61, no. 1, pp. 289–294, 2012.

[12] H. Leder, B. Belke, A. Oeberst, and D. Augustin, "A model of aesthetic appreciation and aesthetic judgments," *Brit. J. Psychol*., vol. 95, no. 4, pp. 489–508, 2004.

[13] A. Chatterjee, "Prospects for a cognitive neuroscience of visual aesthetics," *Bulletin Psychol. and the Arts*, vol. 4, no. 2, pp. 56–60, 2004.

[14] M. W. Greenlee and U. T. Peter, "Functional neuroanatomy of the human visual system: A review of functional MRI studies," in *Pediatric Ophthalmology, Neuro-Ophthalmology, Genetics*. Berlin: Springer, 2008, pp. 119–138.

[15] B. Wandell, S. Dumoulin, and A. Brewer, "Visual cortex in humans," *Encyclopedia of Neuroscience*, vol. 10, pp. 251–257, 2009.

[16] S. Zeki and J. Nash, *Inner Vision: An Exploration of Art and the Brain*. London: Oxford Univ. Press, 1999.

[17] P. Cavanagh, "The artist as neuroscientist," *Nature*, vol. 434, no. 7031, pp. 301–307, 2005.

[18] T. Ang, *Digital Photographer's Handbook*. London: Dorling Kindersley Publishing, 2002.

[19] M. Freeman, *The Photographer's Eye: Composition and Design for Better Digital Photos*. Boca Raton, FL: CRC, 2007.

[20] H. Tong, M. Li, H.-J. Zhang, J. He, and, and C. Zhang, "Classification of digital photos taken by photographers or home users," in *Advances in Multimedia Information Processing*, K. Aizawa, Y. Nakamura, and S. Satoh, Eds., *Lecture Notes in Computer Science*, vol. 3331. Berlin: Springer, 2004, pp. 198–205.

[21] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. European Conf. Computer Vision (ECCV)*. Berlin: Springer, 2006, pp. 288–301.

[22] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or, "Optimizing photo composition," *Computer Graphics Forum*, vol. 29, no. 2, pp. 469–478, 2010.

[23] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "RAPID: Rating pictorial aesthetics using deep learning," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 457–466.

[24] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015, pp. 990–998.

[25] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in Proc. *European Conf. Computer Vision (ECCV)*, 2016, pp. 662–679.

[26] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 497–506.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

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

[29] A. Ebrahimi Moghadam, P. Mohammadi, and S. Shirani, "Subjective and objective quality assessment of image: A survey," *Majlesi J. Elect. Eng.*, vol. 9, Mar. 2015.

[30] A. G. George and K. Prabavathy, "A survey on different approaches used in image quality assessment," *Int. J. Computer Sci. and Network Security*, vol. 14, no. 2, p. 78, 2014.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[32] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *Proc. 1st Int. Workshop Video Processing and Quality Metrics for Consumer Electronics*, 2005, pp. 23–25.

[33] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems 2*, D. S. Touretzsky, Ed. Neural Information Processing Systems Foundation, 1989.

[34] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 38, no. 2, pp. 295–307, 2016.

[35] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," in ACM *Trans. Graphics*, vol. 27, no. 3, article no. 73, 2008.

[36] C. Dong, Y. Deng, C. Change Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015, pp. 576–584.

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

[38] X. Sun, H. Yao, R. Ji, and S. Liu, "Photo assessment based on computational visual attention model," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 541–544.

[39] J. You, A. Perkis, M. M. Hannuksela, and, and M. Gabbouj, "Perceptual quality assessment based on visual attention analysis," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2009, pp. 561–564.

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

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

[42] L.-Y. Lo and J.-C. Chen, "A statistic approach for photo quality assessment," in *Proc. IEEE Int. Conf. Information Security and Intelligence Control*, 2012, pp. 107–110.

[43] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 271–280.

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

[45] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[46] M.-C. Yeh and Y.-C. Cheng, "Relative features for photo quality assessment," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2012, pp. 2861–2864.

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

[48] L. Marchesotti, F. Perronnin, and F. Meylan, "Learning beautiful (and ugly) attributes," in *Proc. British Machine Vision Conf. (BMVC)*, vol. 7, 2013, pp. 1–11.

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

[50] Z. Dong, X. Shen, H. Li, and X. Tian, "Photo quality assessment with DCNN that understands image well," in *Proc. Int. Conf. Multimedia Modeling*, 2015, pp. 524–535.

[51] H. Lv and X. Tian, "Learning relative aesthetic quality with a pairwise approach," in *Proc. Int. Conf. Multimedia Modeling*, 2016, pp. 493–504.

[52] K.-C. Peng and T. Chen, "Toward correlating and solving abstract tasks using convolutional neural networks," in *Proc. IEEE Winter Conf. Applications Computer Vision (WACV)*, 2016, pp. 1–9.

[53] W. Wang, M. Zhao, L. Wang, J. Huang, C. Cai, and X. Xu, "A multi-scene deep learning model for image aesthetic evaluation," *Signal Process.: Image Commun*., vol. 47, pp. 511–518, Sept. 2016.

[54] X. Tian, Z. Dong, K. Yang, and T. Mei, "Query-dependent aesthetic model with deep learning for photo quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2035–2048, 2015.

[55] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.

[56] Z. Wang, F. Dolcos, D. Beck, S. Chang, and T. S. Huang, "Brain-inspired deep networks for image aesthetics assessment," arXiv preprint arXiv:1601.04155, 2016.

[57] L. Zhang, "Describing human aesthetic perception by deeply-learned attributes from Flickr," arXiv preprint arXiv:1605.07699, 2016.

[58] Y. Kao, K. Huang, and S. Maybank, "Hierarchical aesthetic quality assessment using deep convolutional neural networks," *Signal Process.: Image Commun*., vol. 47, pp. 500–510, Sept. 2016.

[59] Y. Kao, R. He, and K. Huang, "Visual aesthetic quality assessment with multitask deep learning," arXiv preprint arXiv:1604.04970, 2016.

[60] R. Datta, J. Li, and J. Z. Wang, "Algorithmic inferencing of aesthetics and emotion in natural images: An exposition," in *Proc. IEEE Int. Conf. Image Processing*, 2008, pp. 105–108.

[61] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Proc. IEEE Int. Conf. Computer Vision (ICCV),* 2011, pp. 2206–2213.

[62] T. O. Ayd in, A. Smolic, and M. Gross, "Automated aesthetic analysis of photographic images," *IEEE Trans. Vis. Comput. Graphics*, vol. 21, no. 1, pp. 31–42, 2015.

[63] L.-K. Wong and K.-L. Low, "Saliency-enhanced image aesthetics class prediction," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2009, pp. 997–1000.

[64] S. Bhattacharya, R. Sukthankar, and M. Shah, "A holistic approach to aesthetic enhancement of photographs," *ACM Trans. Multimedia Computing, Commun., and Applicat.*, vol. 7S, no. 1, 2011.

[65] Y. Wu, C. Bauckhage, and C. Thurau, "The good, the bad, and the ugly: Predicting aesthetic image labels," in *Proc. IEEE Int. Conf. Pattern Recognition (ICPR)*, 2010, pp. 1586–1589.

[66] K.-Y. Lo, K.-H. Liu, and C.-S. Chen, "Assessment of photo aesthetics with efficiency," in *Proc. IEEE Int. Conf. Pattern Recognition (ICPR)*, 2012, pp. 2186–2189.

[67] L. Zhang, Y. Gao, R. Zimmermann, Q. Tian, and X. Li, "Fusion of multichannel local and global structural cues for photo aesthetics evaluation," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1419–1429, 2014.

[68] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. European Conf. Computer Vision (ECCV) Workshop on Statistical Learning in Computer Vision*, Prague, Czech Republic, 2004, pp. 1–2.

[69] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 2007, pp. 1–8.

[70] C. Li, A. Gallagher, A. C. Loui, and T. Chen, "Aesthetic quality assessment of consumer photos with faces," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2010, pp. 3221–3224.

[71] A. Lienhard, P. Ladret, and A. Caplier, "Low level features for quality assessment of facial images," in *Proc. Int. Conf. Computer Vision Theory and Applications (VISAPP*), 2015, pp. 545–552.

[72] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien, "Scenic photo quality assessment with bag of aesthetics-preserving features," in *Proc. ACM Int. Conf. Multimedia*, 2011, pp. 1213–1216.

[73] W. Yin, T. Mei, and C. W. Chen, "Assessing photo quality with geo-context and crowdsourced photos," in *Proc. IEEE Visual Communications and Image Processing Conf.*, 2012, pp. 1–6.

[74] R. Sun, Z. Lian, Y. Tang, and J. Xiao, "Aesthetic visual quality evaluation of Chinese handwritings," in *Proc. Int. Conf. Artificial Intelligence*, 2015, pp. 2510–2516.

[75] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges. L. Bottou, and K. Q. Weinberger, Eds. Neural Information Processing Systems Foundation, 2012, pp. 1097–1105.

[76] T. Joachims, "Training linear SVMs in linear time," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2006, pp. 217–226.

[77] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Int. Conf. Artificial Intelligence and Statistics*, Feb 2015, pp. 562–570.

[78] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proc. IEEE Int. Conf. Pattern Recognition (ICPR)*, 2010, pp. 3121–3124.

[79] C. Huang, Y. Li, C. Change Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 2016, pp. 5375–5384.

[80] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[81] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," arXiv preprint arXiv:1408.5093, 2014.

[82] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Machine Learning Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[83] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Neural Information Processing Systems Foundation, 2014, pp. 3320–3328.

[84] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 2016, pp. 770–778.

[85] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," arXiv preprint arXiv:1602.07261, 2016.

[86] X. Zhang, Z. Li, C. C. Loy, and D. Lin, "Polynet: A pursuit of structural diversity in very deep networks," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, in press.

[87] M. Li, T. Zhang, Y. Chen, and A. J. Smola, "Efficient mini-batch training for stochastic optimization," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2014, pp. 661–670.

[88] D. R. Wilson and T. R. Martinez, "The general inefficiency of batch training for gradient descent learning," *Neural Networks*, vol. 16, no. 10, pp. 1429–1451, 2003.

[89] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 1735–1742.

[90] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Neural Information Processing Systems Foundation, 2005, pp. 1473–1480.

[91] B. Seguin, C. Striolo, F. Kaplan, et al. "Visual link retrieval in a database of paintings," in *Proc. European Conf. Computer Vision (ECCV)*, 2016, pp. 753–767.

[92] Y. Wang and W. Deng, "Self-restraint object recognition by model based CNN learning," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2016, pp. 654–658.

[93] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1335–1344.

[94] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 918–930, 2016.

[95] Y. Kao, C. Wang, and K. Huang, "Visual aesthetic quality assessment with a regression model," in *Proc. Int. Conf. Image Processing*, 2015, pp. 1583–1587.

[96] E. Mavridaki and V. Mezaris, "A comprehensive aesthetic quality assessment method for natural images using basic rules of photography," in *Proc. Int. Conf. Image Processing*, 2015, pp. 887–891.

[97] Z. Dong and X. Tian, "Multi-level photo quality assessment with multi-view features," *Neurocomputing*, vol. 168, pp. 308–319, Nov. 2015.

[98] G.-S. Xie, X.-Y. Zhang, S. Yan, and C.-L. Liu, "Hybrid CNN and dictionary-based models for scene recognition and domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, 2015.

[99] N. Jaiswal and Y. K. Meghrajani, "Saliency based automatic image cropping using support vector machine classifier," in *Proc. Int. Conf. Innovations Information, Embedded and Communication Systems*, 2015, pp. 1–5.

[100] J. Sun and H. Ling, "Scale and object aware image thumbnailing," *Int. J. Comput. Vision*, vol. 104, no. 2, pp. 135–153, 2013.

[101] E. Ardizzone, A. Bruno, and G. Mazzola, "Saliency based image cropping," in *Proc. Int. Conf. Image Analysis and Processing*, 2013, pp. 773–782.

[102] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato, "Sensation-based photo cropping," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 669–672.

[103] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 291–300.

[104] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, and C. Chen, "Probabilistic graphlet transfer for photo cropping," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 802–815, 2013.

[105] J. Yan, S. Lin, S. B. Kang, and X. Tang, "Change-based image cropping with exclusion and compositional features," *Int. J. Comput. Vision*, vol. 114, no. 1, pp. 74–87, 2015.

[106] J. Yan, S. Lin, S. Kang, and X. Tang, "Learning the change for automatic image cropping," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 2013, pp. 971–978.

[107] J. Huang, H. Chen, B. Wang, and S. Lin, "Automatic thumbnail generation based on visual representativeness and foreground recognizability," in *Proc. IEEE Int. Conf. Computer Vision (ICCV),* 2015, pp. 253–261.

[108] J. Chen, G. Bai, S. Liang, and Z. Li, "Automatic image cropping: A computational complexity study," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 507–515.

[109] J. Park, J.-Y. Lee, Y.-W. Tai, and I. S. Kweon, "Modeling photo composition and its application to photo re-arrangement," in *Proc. Int. Conf. Image Processing*, 2012, pp. 2741–2744.

[110] P. Wang, Z. Lin, and R. Mech, "Learning an aesthetic photo cropping cascade," in *Proc. IEEE Winter Conf. Applications Computer Vision (WACV),* 2015, pp. 448–455.

[111] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," arXiv preprint arXiv:1312.6229, 2013.

SP