arXiv:1604.04970v2 [cs.CV] 20 Aug 2016

Deep Aesthetic Quality Assessment with Semantic Information

Yueying Kao, Ran He, Kaiqi Huang, Tieniu Tan

***Abstract*—Human beings often assess the aesthetic quality of an image coupled with the identification of the image's semantic content. This paper addresses the correlation issue between automatic aesthetic quality assessment and semantic recognition. We cast the assessment problem as the main task among a multi-task deep model, and argue that semantic recognition task offers the key to address this problem. Based on convolutional neural networks, we employ a single and simple multi-task framework to efficiently utilize the supervision of aesthetic and semantic labels. A correlation item between these two tasks is further introduced to the framework by incorporating the inter-task relationship learning. This item not only provides some useful insight about the correlation but also improves assessment accuracy of the aesthetic task. Particularly, an effective strategy is developed to keep a balance between the two tasks, which facilitates to optimize the parameters of the framework. Extensive experiments on the challenging AVA dataset and Photo.net dataset validate the importance of semantic recognition in aesthetic quality assessment, and demonstrate that multi-task deep models can discover an effective aesthetic representation to achieve state-of-the-art results.**

***Index Terms*—Visual aesthetic quality assessment, semantic information, multi-task learning.**

I. Introduction

Aesthetic image analysis has attracted increasing attention in computer vision community [1], [2], [3], [4], [5], [6], [7], [8]. It is related to the high-level perception of visual aesthetics. Machine learning models for visual aesthetic quality assessment have shown to be useful in many applications, e.g., image retrieval, photo management, image editing, and photography [9], [10], [11], [12]. Since visual aesthetics is a subjective attribute, automatically assessing aesthetic quality of images is still challenging. Many data-driven approaches [13], [14], [15], [16], [17], [3], [18], [19], [20], [21] have been proposed to address this issue. These methods often learn from the aesthetic quality of images that are labeled by humans. Most of these methods aim to discover a meaningful and better aesthetic representation, and often formulate the representation learning as a single and standalone classification task.

Handcrafted features are earlier attempts. They are based on the intuitions of how people perceive the aesthetic quality of images or photographic rules. These features include color [10], [13], [22], the rule of thirds [13], simplicity [14], [3], and composition [15]. Later, generic image descriptors such as bag-of-visual-words (BOV) [23] and fisher vectors (FV) [24] are used to assess aesthetic quality. They are shown to outperform the traditional handcrafted features [16], [25], [26]. Recently, deep convolutional neural networks (CNNs) [27], [28] have been applied to aesthetic quality assessment [29], [30], [31], [32]. Nevertheless, these computational approaches provide either accurate or interpretable results [4].

For human beings, aesthetic quality assessment is always coupled with the identification of semantic content of images [33], [34]. It is difficult for humans to treat aesthetic quality assessment as an isolate and independent task. When humans assess the aesthetic quality of an image, they first understand what they are assessing. That is, they have known the sematic information of this image. Seen from Fig. 1, we can recognize the semantic content from these images at a glance and assess the aesthetic quality quickly. Hence it is reasonable to assume that, assessing aesthetic quality and semantic recognition are correlated tasks for machine learning. However, the relationship between semantic recognition and automatically assessing visual aesthetic quality has not been fully explored.

This paper addresses the correlation issue between automatic aesthetic quality assessment and semantic recognition. We employ multi-task convolutional neural network to explore the potential correlation. Multi-task learning can learn multiple related tasks in parallel with shared knowledge. It has been demonstrated that this approach can boost some or all of the tasks [35]. Our goal is to utilize semantic recognition in the joint objective function to improve the aesthetic quality assessment, our main task. However, there is still a typical challenge in the multi-task learning for our multi-task problem. That is, the aesthetic task and semantic task face the different learning difficulties. The main reason is that the semantic recognition is much easier than aesthetics assessment. The semantic content is much objective, while the aesthetic attributes are subjective. Thus, different from the strategies of treating all tasks equally and early stopping [35], [36], [37] we present a strategy to keep the effect of both tasks balanced in the joint objective function.

In addition, to discover the relationships between aesthetic and semantic tasks automatically and to better exploit the inter-task relatedness for more effective feature learning, we model the task relationship and impose it in the objective function. To some extent, it can explain the factors in aesthetic quality assessment and make our results more interpretable. Thus, to investigate how to make full use of semantic information and

Yueying Kao, Ran He, Kaiqi Huang, and Tieniu Tan are with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing, 100049, China. Ran He, Kaiqi Huang, and Tieniu Tan are also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China (e-mail: yueying.kao@nlpr.ia.ac.cn; rhe@nlpr.ia.ac.cn; kqhuang@nlpr.ia.ac.cn; tnt@nlpr.ia.ac.cn).

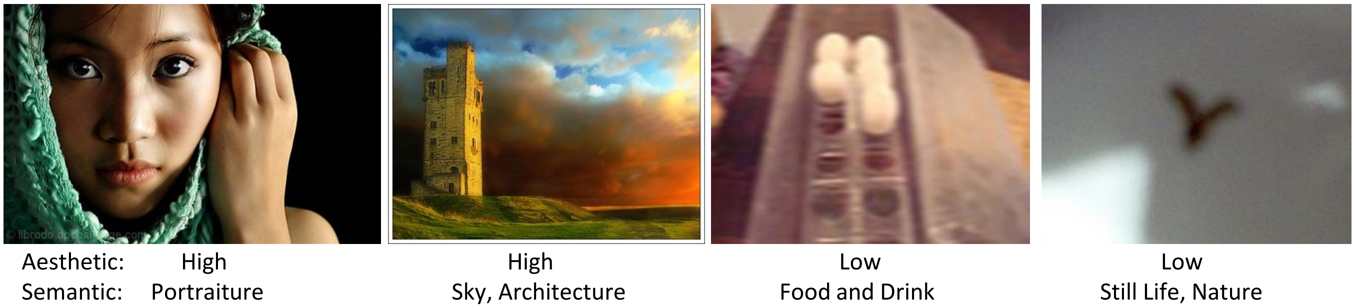


Fig. 1. Example images with their aesthetic and semantic labels on AVA dataset.

how semantic information influence aesthetic task, our multi-task framework considers the strategy of keeping the effect of two tasks balanced and the relationship learning between semantic and aesthetic tasks.

In the evaluation, the most challenging large-scale AVA dataset [25] is used to verify the effectiveness of semantic information for aesthetic feature learning and investigate the correlation among aesthetic and semantic content recognitions. The experiments show that our results significantly outperform the state-of-the-art results [29], [31], [32] for aesthetic quality assessment on AVA dataset. Furthermore, it is demonstrated that the learned representation with our multi-task framework can be transferred for the dataset (here we use Photo.net dataset [1], [13]) with only aesthetic labels and other semantic representation (such as from Imagenet) can also be used for aesthetic representation learning.

Our contributions lie in three-fold:

- Instead of taking visual aesthetic quality assessment as an isolated task, we propose to exploit the semantic recognition to jointly assess the aesthetic quality with a single multi-task convolutional neural network (MTCNN). It is a novel attempt to learn aesthetic features with the help of a related task, i.e. semantic recognition.
- We propose to automatically learn the correlations between the aesthetic and semantic tasks by simultaneously modeling the inter-task relationship and controlling the parameters' complexity of each task in our multi-task framework. It can explain the factors in aesthetic quality assessment and makes our results more interpretable.
- Facing the different learning difficulties between the two tasks, we present a strategy to keep the effect of both tasks balanced in the joint objective function. The proposed method outperforms the state-of-the-art methods on the challenging AVA dataset and Photo.net dataset.

The rest of this paper is organized as follows: we summarize related work in Section II, describe our method in detail in Section III, present the experiments in Section IV, and conclude the paper in Section V.

II. Related work

Since our work is related to the aesthetic quality assessment and multi-task learning, we will mainly review work related to the two parts in this section.

A. *Aesthetic quality assessment*

Most previous works [13], [10], [15], [16], [38], [39] on aesthetic quality assessment focus on the challenging problem of designing appropriate features. Typically, handcrafted features are proposed based on the intuitions about human perception of the aesthetic quality of images or photographic rules. For example, Datta et al. [13] design certain visual features such as colorfulness, the rule of thirds, and low depth of field indicators, to discriminate between aesthetically pleasing and displeasing images. Dhar et al. [15] extract some high level attributes including compositional, content, and sky-illumination attributes, which are characteristically used by humans to describe images. Luo et al. [38] and Tang et al. [3] consider that photos may have different aesthetic criteria in mind for different type of images and design visual features in different ways according to the variety of photo content. In [16], generic image descriptors are used to assess aesthetic quality, which are shown to outperform the traditional handcrafted features.

Despite the success of handcrafted features and generic image descriptors, CNNs have been applied to aesthetic quality assessment [29], [30], [31], [32] and obtain the state-of-the-art performance. CNNs learn aesthetic features automatically. However, they extract features by treating aesthetic quality assessment as an independent problem. The network in [29], RDCNN, hopes to leverage the idea of multi-task learning with the style attributes to help determine the aesthetic quality of images. Unfortunately, due to many missing labels for style attributes, they can not jointly perform aesthetics categorization and style classification in a neural network, and just concatenate the features of the aesthetics and style by using transfer learning. Our work is also related to CNNs for aesthetics classification. In contrast, firstly, we exploit semantic information to assist in learning aesthetic representation with a multi-task learning framework. We can jointly learn aesthetics categorization and semantic recognition with a single multi-task network, which is different from RDCNN [29]. Secondly, our multi-task CNN considers the strategy of keeping the effect of two tasks balanced and the relationship learning between semantic and aesthetic tasks. Finally, images are labeled with semantic information much easier than style attributes in real world. This is because only professional photographer and photography amateurs are familiar with all the style attributes.

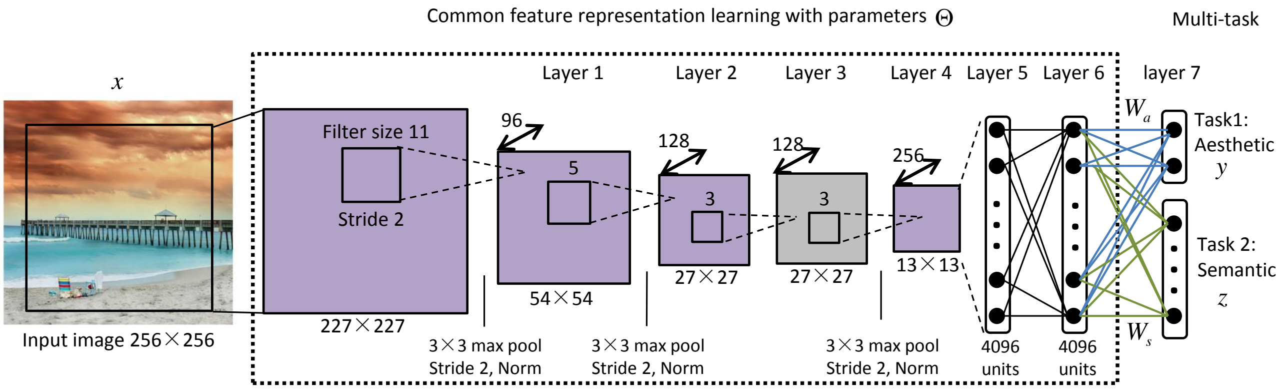


Fig. 2. An illustration for the architecture of our MTCNN #1.

B. Multi-task learning

Multi-task learning aims to boost the generalization performance by learning multiple related tasks simultaneously [35], [40], [37], [41]. It does this by learning tasks in parallel while using a shared representation [35]. Deep neural network can learn features jointly under multiple objectives and it is the earliest models for multi-task learning. Multi-task learning based on deep neural network has been applied to many computer vision problems [37], [36], [42]. However, there are many strategies for sharing knowledge and learning process for different problems. For example, Zhang et al. [43] share parameters in all layers and learn the common features for all tasks, while Liu et al. [44] just sharing in some bottom layers and learn respective representation in some top layers for each task. Yim et al. [36] treat all tasks equally important. In contrast, early stopping strategy is used in some related tasks [37], due to different learning difficulties and convergence rates in different tasks. In our problem, because semantic recognition task is much easier than aesthetic quality assessment, common features of our two tasks are learned simultaneously and an effective strategy of keeping effect of all the tasks balanced in the joint objective function is used. In addition, the task relationships can be learned from the data automatically in the conventional methods [45], [46], [47]. Inspired by this, we consider the relationship learning in our multi-task neural networks to explore the relationships between the aesthetic and semantic tasks.

III. Method

In this section, we propose to exploit the semantic information to help identify the aesthetic quality of images, assuming that they are considered as the related attributes [33], [34]. Here the aesthetic quality assessment is our main task and the semantic content recognition is the aided task. Our problem is firstly formulated as a multi-task convolutional neural network (MTCNN) model without learning task relationships automatically from data. Then we develop a multi-task relationship learning convolutional neural network (MTRLCNN) model by adding the task relationship learning in the objective function to discover the correlation between aesthetic task and semantic tasks. An example of MTCNN architectures is illustrated in Fig. 2. Furthermore, we explore and adapt different network structures to our problem.

A. Multi-Task Probabilistic Framework

Our problem can be interpreted as a probabilistic model. Using the probabilistic formulation, various deep networks can solve our problem by optimizing the model parameters that maximize the posterior probability. Then, Bayesian analysis is leveraged to predict most likely aesthetic quality and semantic attributes of given images.

Assuming a training dataset with a total of N samples, which are associated with C aesthetic classes and M semantic attributes. Considering each image has only one aesthetic class and multiple semantic attributes in real world, each image is represented as $(x_n, y_n, z_n), n = 1, 2, ..., N$. Here x_n represents the n-th image sample, $y_n = c, c = 0, ..., C-1$ is the aesthetic label and $z_n = [z_n^1, ..., z_n^m, ..., z_n^M]^{\mathrm{T}}$ is the semantic label for the n-th image sample. If the n-th image sample has the m-th semantic attribute, the m-th semantic label is set as $z_n^m = 1$, otherwise $z_n^m = 0$. Therefore a given dataset is denoted as $(X, Y, Z) = \{(x_n, y_n, z_n), n \in \{1, 2, ..., N\}\}$. For our MTCNNs (our MTCNN #1 is shown in Fig. 2), Θ denotes the common parameters in some bottom layers to learn features for all tasks, and $W = [W_a, W_s]$ indicates the specific parameters for associated tasks. W_a and W_s represent the parameters for aesthetic quality assessment and semantic recognition respectively. Each column in W_a or W_s corresponds to a subtask. The goal is to find the optimal or sub-optimal parameters Θ, W, λ by maximizing the following posterior probability

$$\hat{\Theta}, \hat{W}, \hat{\lambda} = \underset{\Theta, W, \lambda}{\operatorname{argmax}}\, p(\Theta, W, \lambda | X, Y, Z), \tag{1}$$

where λ is the weight coefficient of the semantic recognition task in the joint learning process.

Based on the Bayesian theorem, we have

$$\begin{aligned} p(\Theta, W, \lambda | X, Y, Z) &= \frac{p(X, Y, Z | \Theta, W, \lambda) p(\Theta, W, \lambda)}{p(X, Y, Z)} \\ &\propto p(X, Y, Z | \Theta, W, \lambda) p(\Theta, W, \lambda), \end{aligned} \tag{2}$$

where $p(X, Y, Z | \Theta, W, \lambda)$ is the conditional probability, and $p(\Theta, W, \lambda)$ is the prior probability.

Then Eqn. (1) takes the form

$$\begin{aligned}&\hat{\Theta}, \hat{W}, \hat{\lambda}\\ &\propto \underset{\Theta, W, \lambda}{\operatorname{argmax}}\, p(Y|X, \Theta, W_a)p(Z|X, \Theta, W_s, \lambda)p(\Theta)p(W)p(\lambda).\end{aligned} \tag{3}$$

Each term in Eqn. (3) is defined as:

1) The conditional probability $p(Y|X, \Theta, W_a)$ corresponds to the task of aesthetic quality assessment. Here assessing aesthetic quality is interpreted as a classification problem and modeled as a multinomial logistic regression similar to traditional classification problems [27]. The conditional probability $p(Y|X, \Theta, W_a)$ can be formulated as

$$p(Y|X, \Theta, W_a) = \prod_{n=1}^{N} \sum_{c=1}^{C} 1\{y_n = c\} p(y_n = c|x_n, \Theta, W_a), \tag{4}$$

where $1\{\cdot\}$ is the indicator function, it has two values, $1\{a\ true\ statement\} = 1$, and $1\{a\ false\ statement\} = 0$. $p(y_n = c|x_n, \Theta, W_a)$ is calculated by the softmax function

$$p(y_n = c|x_n, \Theta, W_a) = \frac{\exp({W_a^c}^{\mathrm{T}}(\Theta^{\mathrm{T}} x_n))}{\sum_{l=1}^{C} \exp({W_a^l}^{\mathrm{T}}(\Theta^{\mathrm{T}} x_n))}. \tag{5}$$

2) The conditional probability $p(Z|X, \Theta, W_s, \lambda)$ corresponds to the semantic recognition. Since each element of the semantic label of a given image is binary: $z_n^m \in \{0, 1\}$, each semantic attribute recognition can be interpreted as a logistic regression. Hence the conditional probability $p(Z|X, \Theta, W_s, \lambda)$ can be

$$\begin{aligned}&p(Z|X, \Theta, W_s, \lambda)\\ &= \prod_{n=1}^{N} \prod_{m=1}^{M} (p(z_n^m = 1|x_n, \Theta, W_s^m)^{z_n^m}\\ &(1 - p(z_n^m = 1|x_n, \Theta, W_s^m))^{1-z_n^m})^{\lambda},\end{aligned} \tag{6}$$

where $p(z_n^m = 1|x_n, \Theta, W_s^m)$ is calculated by a sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$.

3) The prior probability $p(\Theta)$ corresponds to the network parameters for common features. The parameters Θ can be initialized as a standard normal distribution like previous network [27]. $p(\Theta) = \prod_{k=1}^{K} p(\theta_k) = \prod_{k=1}^{K} N(\mathbf{0}, I)$, where $\mathbf{0}$ is a zero matrix and I is an identity matrix.

4) Similar to Θ, the parameters W for specific tasks can also be initialized as a standard normal distribution. Thus, the prior probability can be $p(W) = p(W_a)p(W_s) = N_a(\mathbf{0}, I)N_s(\mathbf{0}, I)$.

5) λ is used to control the influence of semantic recognition task in the final objective function. The prior probability $p(\lambda)$ is implemented by defining λ obeying a normal distribution, $p(\lambda) = N(\mu, \sigma^2)$.

Then Eqns. (4), (5) and (6) are substituted into Eqn. (3), negative log function is taken for Eqn. (3), and the constant terms are omitted. As a result, the objective function can be

$$\begin{aligned}&\underset{\Theta, W, \lambda}{\operatorname{argmin}}\{-\sum_{n=1}^{N}\sum_{c=1}^{C} 1\{y_n = c\} log\frac{\exp({W_a^c}^{\mathrm{T}}(\Theta^{\mathrm{T}} x_n))}{\sum_{l=1}^{C} \exp({W_a^l}^{\mathrm{T}}(\Theta^{\mathrm{T}} x_n))}\\ &-\lambda\sum_{n=1}^{N}\sum_{m=1}^{M}(z_n^m log\sigma({W_s^m}^{\mathrm{T}}(\Theta^{\mathrm{T}} x_n)) + (1 - z_n^m)(1-\\ &log\sigma({W_s^m}^{\mathrm{T}}(\Theta^{\mathrm{T}} x_n)))) + \Theta^{\mathrm{T}}\Theta + W^{\mathrm{T}}W + (\lambda - \mu)^2\}.\end{aligned} \tag{7}$$

B. Multi-Task Relationship Learning Probabilistic Framework

To automatically learn the relationships between aesthetic and semantic tasks and to better exploit the inter-task relatedness for aesthetic feature learning, we model the relationships between tasks as a covariance matrix Ω and add it to our above multi-task framework. The new framework is called Multi-Task Relationship Learning (MTRL) framework. In the MTRL framework, the goal is to find the optimal or sub-optimal parameters $\Theta, W, \lambda, \Omega$ by maximizing the following posterior probability

$$\hat{\Theta}, \hat{W}, \hat{\lambda} = \underset{\Theta, W, \lambda}{\operatorname{argmax}}\, p(\Theta, W, \lambda, \Omega|X, Y, Z), \tag{8}$$

Based on the Bayesian theorem, Eqn. (8) takes the form

$$\begin{aligned}\hat{\Theta}, \hat{W}, \hat{\lambda} \propto \underset{\Theta, W, \lambda}{\operatorname{argmax}} p(Y|X, \Theta, W_a)p(Z|X, \Theta, W_s, \lambda)\cdot\\ p(W|\Omega)p(\Theta)p(W)p(\lambda).\end{aligned} \tag{9}$$

The conditional probability $p(Y|X, \Theta, W_a)$, the conditional probability $p(Z|X, \Theta, W_s, \lambda)$, the prior probability $p(\Theta)$, the prior probability $p(W)$ and the prior probability $p(\lambda)$ are same to the above definition in Section III-A. For the prior on the W, we consider two terms $p(W)$ and $p(W|\Omega)$. The prior probability $p(W)$ is to model the each column of W as a standard normal distribution for each task and can separately penalize the complexity of the each column of W. The $p(W|\Omega)$ is to model the structure of W between tasks by using a matrix-variate normal distribution [45], [48]. So we have

$$\begin{aligned}p(W|\Omega) &= MN(0, I \otimes \Omega)\\ &= \frac{exp(-\frac{1}{2}tr(I^{-1}W\Omega^{-1}W^T))}{(2\pi)^{d(M+C)/2}|I|^{(M+C)/2}|\Omega|^{d/2}},\end{aligned} \tag{10}$$

where d is the dimension of the common representation of all the tasks, such as the dimension of layer 7 in Fig. 2. The new objective function can be

$$\begin{aligned}&\underset{\Theta, W, \lambda}{\operatorname{argmin}}\{-\sum_{n=1}^{N}\sum_{c=1}^{C} 1\{y_n = c\} log\frac{\exp({W_a^c}^{\mathrm{T}}(\Theta^{\mathrm{T}} x_n))}{\sum_{l=1}^{C} \exp({W_a^l}^{\mathrm{T}}(\Theta^{\mathrm{T}} x_n))}\\ &\quad-\lambda\sum_{n=1}^{N}\sum_{m=1}^{M}(z_n^m log\sigma({W_s^m}^{\mathrm{T}}(\Theta^{\mathrm{T}} x_n)) + (1 - z_n^m)(1-\\ &\quad log\sigma({W_s^m}^{\mathrm{T}}(\Theta^{\mathrm{T}} x_n)))) + \Theta^{\mathrm{T}}\Theta + W^{\mathrm{T}}W + (\lambda - \mu)^2\\ &\quad + tr(W\Omega^{-1}W^T)\},\\ &s.t. \quad \Omega \geq 0, \quad tr(\Omega) = 1.\end{aligned} \tag{11}$$

Where the constraint $tr(\Omega) = 1$ is the same as in [45].

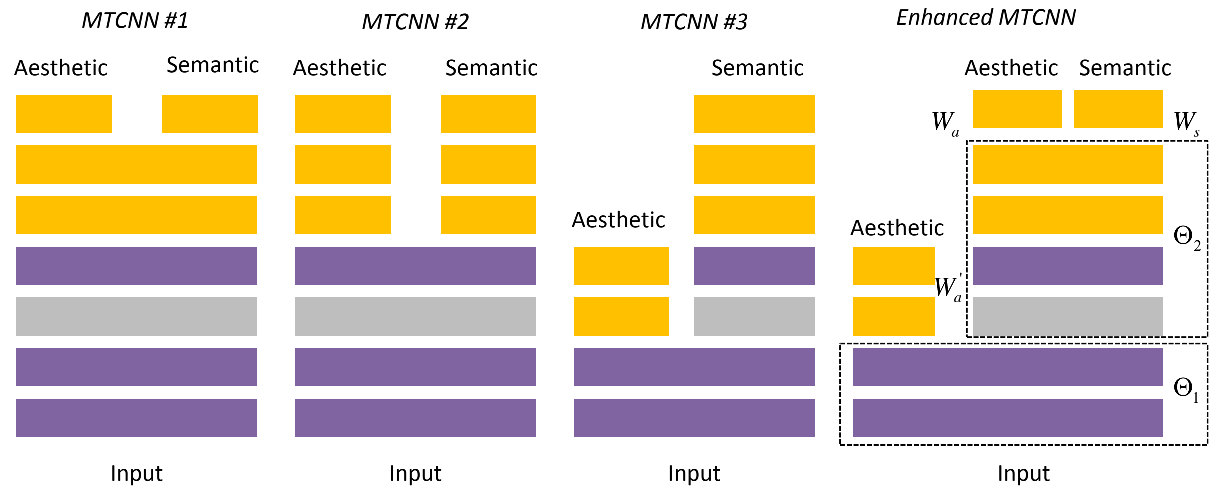


Fig. 3. Explored MTCNNs with different architectures. The details of MTCNN #1 are illustrated in Fig. 2. Color code used: purple = convolutional layer + max pooling, grey = convolutional layer, yellow = fully-connected layer.

C. Optimization Procedure

The multi-task objective function in Eqn. (7) and (11) can be optimized by a network through stochastic gradient descent (SGD) [27]. Here a specific CNN is applied to search optima for the parameters $\Theta, W, \lambda, \Omega$. One architecture of our MTCNNs is shown in Fig. 2. For the optimization procedure of MTCNNs, firstly, all tasks share knowledge in bottom layers. Then specific features are learned for each task in top layers. Finally, the combination of the softmax loss function for aesthetic quality prediction (the first term in Eqn. (7)) and the cross entropy loss function for semantic recognition (the second term in Eqn. (7)) are employed to update the parameters of the network jointly by back propagation. For the MTRLCNN, we adopt an alternate optimization procedure [45] to minimize the objective function in Eqn. (11) for the parameters Θ, W, Ω. Firstly, we update Θ, W by back propagation like the MTCNN with fixed Ω. Then fix Θ, W and optimize the Ω, $\Omega = \frac{(W^T W)^{1/2}}{tr(W^T W)^{1/2})}$. We repeat this procedure until convergence.

Traditionally, multiple tasks are treated equally important in back propagation of multi-task learning [35], [36] assuming that they can reach best performance roughly at the same time. However, different tasks may have different learning difficulties and convergence rates. Caruana [35] propose to control the effect of different tasks by adjusting the learning weight on each output task. He also put forward some strategies for this problem, such as early-stopping. Early stopping strategy has been used to some works [37] and good performance is achieved. Nevertheless, this strategy is not suited to our problem. This is because the extra task (i.e., semantic recognition task) is much easier, and often converges more quickly than the main task (i.e., aesthetic quality assessment). Our experimental results (details in Table I and Section IV) show that, if the convergent semantic recognition task is early stopped, the training loss of the aesthetic task will do not drop obviously and converge in a low rate. We think that it is mainly because the aesthetic is subjective and needs the help of semantic task in entire training process. Hence, we present a simple strategy to keep the effect of all tasks balanced in back propagation. Because the softmax loss function only considers the value corresponding ground truth label for each example. In our problem, $\lambda = 1/M$ is fixed in the objective function in the entire training process.

D. Network Architectures Implementation Exploration

To implement the multi-task model, we investigate several multi-task network architectures to utilize semantic information for visual aesthetic quality assessment. Take the MTCNN as an example and adapt the networks to our problem, then apply suited network architecture to our MTRLCNN. These networks are explained in Fig. 3. The supervision of aesthetic and semantic labels can be in the same or different layers in the network. Here we propose and explore three basic network architectures and an enhanced network. For all networks, the input is a $227 \times 227 \times 3$ patch randomly extracted from a resized image $256 \times 256 \times 3$ as previous work [29].

MTCNN #1: Since our goal is to discover the effective features for aesthetic assessment with the help of semantic information, a simple idea is to learn all parameters for aesthetic representations with aesthetic and semantic supervision in a network until the last layers. MTCNN #1 implements this idea. The architecture of MTCNN #1 (in Fig. 3) is detailed in Fig. 2. The network contains four convolutional layers and two fully-connected layers with parameters Θ for common feature learning. Then the network is split into two branches, the two last layers for two specific tasks. Thus the parameters $W = [W_a, W_s]$ from layer 6 to layer 7 for each task are learned separately. Then, the softmax loss function is adopted for aesthetic quality prediction, and the cross entropy loss function for semantic recognition. The combination of the two loss functions is employed to jointly update the parameters of the network.

MTCNN #2: To explore different structures for aesthetic features learning, we introduce MTCNN #2 (shown in Fig. 3) to allow some top layers to learn aesthetic representations independently without semantic supervision. Similar to MTCNN

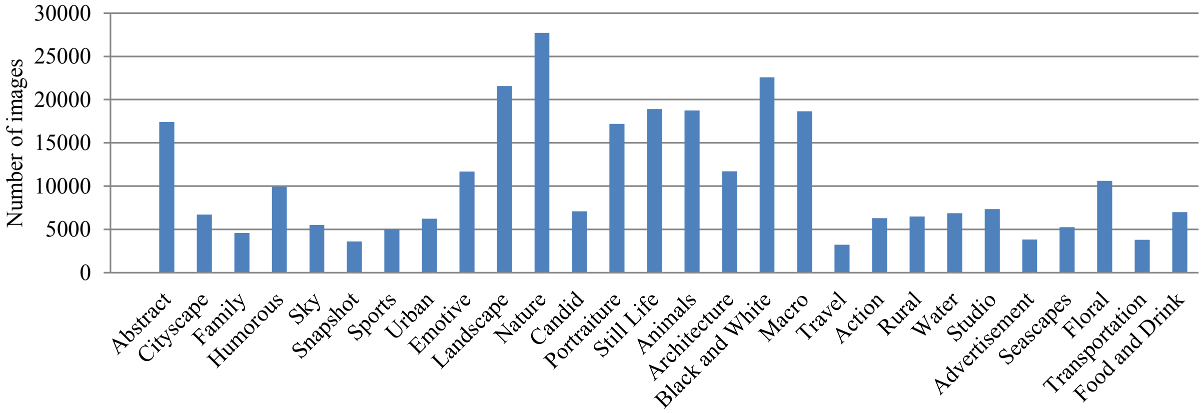


Fig. 4. The number of images for each semantic tag on AVA dataset.

#1, the network #2 contains four convolutional layers with parameters Θ for common feature learning. Then the network is split into two branches earlier than MTCNN #1 for two specific tasks. Different from the architecture #1, layers 5, 6 and 7 in the network #2 learn parameters $W = [W_a, W_s]$ separately for the two tasks. The loss functions are also the same as the architecture #1.

MTCNN #3: Since CNNs can learn hierarchical features, we consider the low-level features of a network for our main task in the MTCNN #3 (shown in Fig. 3). In this network, four convolutional layers and three fully-connected layers are designed for semantic recognition, while two convolutional layers and two fully-connected layers for aesthetic quality assessment. The two tasks share knowledge Θ in the two convolutional layers. The other layers are used to learn specific parameters $W = [W_a, W_s]$ for each task. The loss functions are also the same as the architecture #1.

Enhanced MTCNN: To further explore the effective aesthetic features, we propose an enhanced MTCNN by combining MTCNN #1 and MTCNN #3. That is, we add extra aesthetic supervision in the first two layers in MTCNN #1. Shown in Fig. 3, the common parameters Θ_1 in the first and second convolutional layers are learned for three tasks, the common parameters Θ_2 in other two convolutional layers and two fully-connected layers are learned for two tasks, and specific parameters $W = [W_a^{'}, W_a, W_s]$ are learned separately in top layers. Our goal is to enhance the supervision of aesthetic labels in the first and second convolutional layers under the premise of ensuring the influence of semantic information in all network. Here we denote $\Theta = [\Theta_1, \Theta_2]$. The objective function in Eqn. (7) is transformed to

$$\begin{aligned}&\underset{\Theta,W,\lambda}{\operatorname{argmin}}\{-\sum_{n=1}^{N}\sum_{c=1}^{C}1\{y_n=c\}log\frac{\exp({W_a^{c}}^{\mathrm{T}}(\Theta^{\mathrm{T}}x_n))}{\sum_{l=1}^{C}\exp({W_a^{l}}^{\mathrm{T}}(\Theta^{\mathrm{T}}x_n))}\\&-\sum_{n=1}^{N}\sum_{c=1}^{C}1\{y_n=c\}log\frac{\exp({W_a^{'c}}^{\mathrm{T}}(\Theta_1^{\mathrm{T}}x_n))}{\sum_{l=1}^{C}\exp({W_a^{'l}}^{\mathrm{T}}(\Theta_1^{\mathrm{T}}x_n))}\\&-\lambda\sum_{n=1}^{N}\sum_{m=1}^{M}(z_n^m log\sigma({W_s^{m}}^{\mathrm{T}}(\Theta^{\mathrm{T}}x_n))+(1-z_n^m)(1-\\&log\sigma({W_s^{m}}^{\mathrm{T}}(\Theta^{\mathrm{T}}x_n))))+\Theta^{\mathrm{T}}\Theta+W^{\mathrm{T}}W+(\lambda-\mu)^2\},\end{aligned} \tag{12}$$

where the first term in Eqn. (12) is our main task, and the second term is the added task. We fix $\lambda = 2/M$ based on our strategy for the enhanced MTCNN.

E. Transfer learning with semantic information

Semantic content recognition has been studied for many years in computer vision, such as object recognition, object detection, image classification and semantic segmentation [49], [27], [50], [51], [52]. Recently, deep learning methods have achieved great succuss on the semantic recognition, especially the image classification on Imagenet [27], [51], [52], [53]. The Imagenet [53] dataset contains rich semantic information and can be utilized to further help aesthetic representation learning. Thus we transfer the semantic representation learned from the network pretrained on Imagenet to aesthetic quality assessment. A trained model on a dataset can be transferred to another dataset for a similar or different task [54], [55]. Specifically, our multi-task architecture from Layer 1 to Layer 6 in Fig. 2 is replaced with AlexNet [27], VGG Net [51] or ResNet [52]. It is shown MTCNN #1 performs best in the three basic MTCNNs from Table II. We initialize the networks with models pretrained on Imagenet and finetune it with the training data labeled with aesthetic labels and semantic labels.

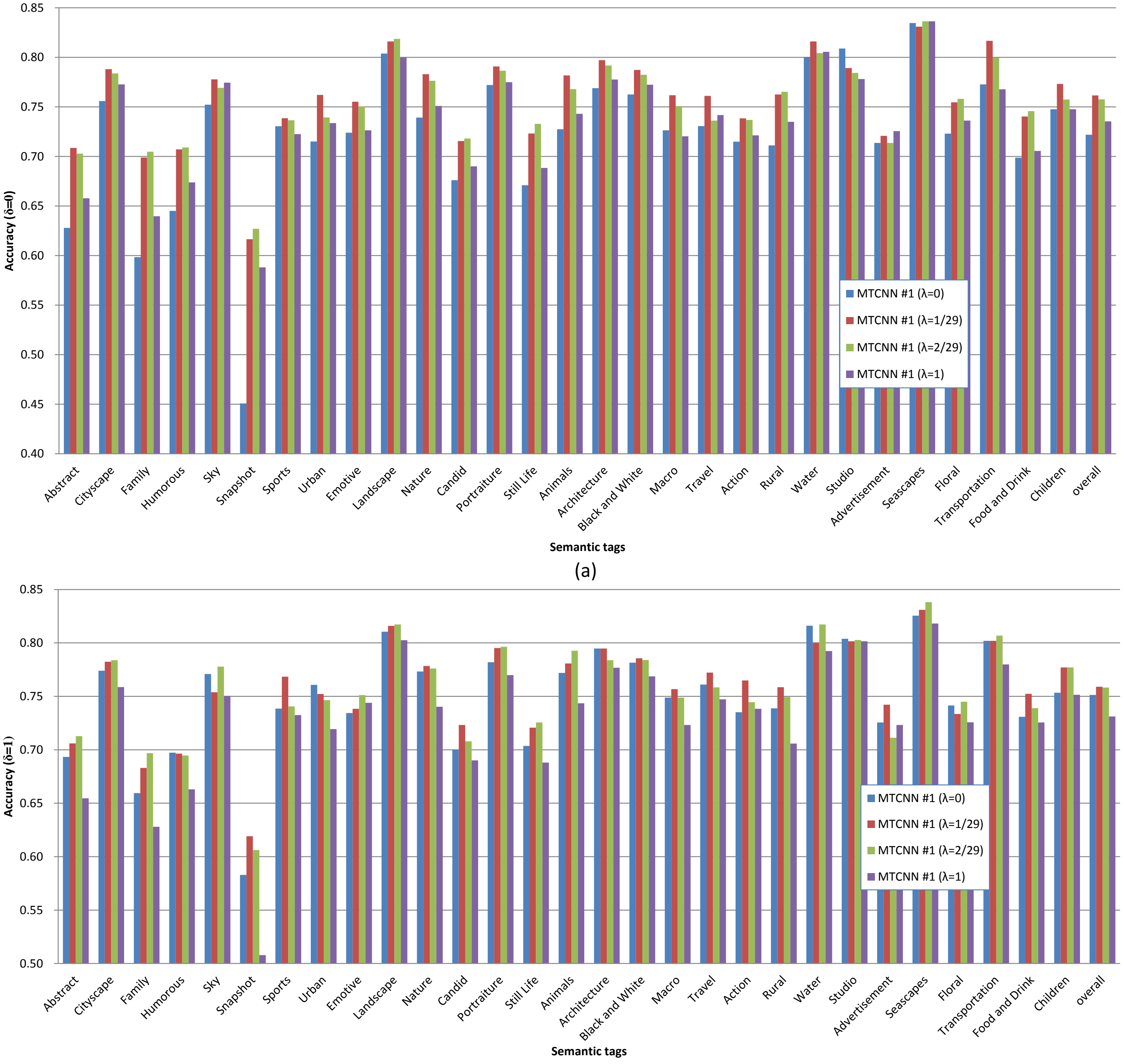
In addition, another meaningful direction is how to exploit the massive dataset of visual semantic understanding for the limited dataset with only aesthetic labels for aesthetic assessment. To transfer the learned representation with both aesthetic and semantic supervision to the dataset with only aesthetic labels, we initialize the networks with pretrained multi-task models and finetune it with the training data labeled with only aesthetic labels.

IV. Experiments

In this section, we evaluate the proposed method on the challenging large-scale AVA dataset and Photo.net dataset. Experimental results show that the benefits of semantic information and the effectiveness of our proposed method.

A. Dataset

AVA dataset: The AVA dataset [25] is one of the most large-scale and challenging dataset for visual aesthetic quality



(a)

(b)

Fig. 5. Accuracy on each semantic tag using MTCNN #1 with different λ when $\delta = 0$ and $\delta = 1$ on the AVA dataset.

assessment. It contains more than 255,000 images gathered from *www.dpchallenge.com*. Each image has about 200 voters to assess the aesthetic score from one to ten. In addition, each image contains 0, 1 or 2 semantic tags (attributes). We select 185,751 images used in this paper based on the following rules. 1) More than 3000 images are available for each tag; 2) each image contains at least one tag. Eventually 29 semantic tags are chosen and the number of images for each tag is listed in Fig. 4. From the 185,751 images, 20,000 images are randomly selected as the testing set similar to [29], and the rest 165,751 images as the training set. For aesthetic labels, we follow the experimental setup as [25], [29], the training set is divided into two classes: high quality and low quality images. We designate the images with an average score larger than $5+\delta$ as high quality images, those with an average score smaller than $5 - \delta$ as low quality images. Images with an average score between $5+\delta$ and $5-\delta$ are discarded. We set δ to 0 and 1 respectively for the training set to obtain the ground truth labels. There are 165,751 images in the training set when $\delta = 0$ and 38,994 images in the training set when $\delta = 1$. We set δ to 0 for the testing set regardless of the value of δ for the training set. For semantic labels, each image is labeled as a 29-dim binary vector.

Photo.net dataset[1]: The Photo.net dataset [1], [13] is a dataset

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

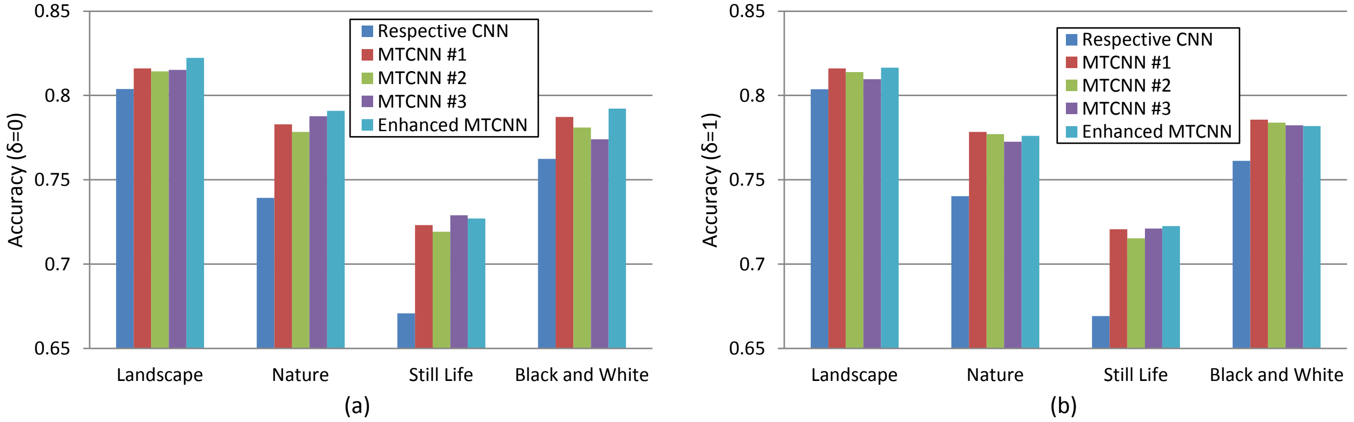


Fig. 6. The accuracy with different methods for aesthetic classification on "Landscape", "Nature", "Still Life" and "Black and White" separately with both $\delta = 0$ and $\delta = 1$.

TABLE I
ACCURACY (%) OF OUR MTCNN #1 WITH DIFFERENT λ ON THE AVA DATASET.

δ	$\lambda = 0$	$\lambda = 1/29$	$\lambda = 2/29$	$\lambda = 1$	with early stopping
0	72.19	**76.15**	75.76	73.54	73.43
1	75.13	**75.90**	75.82	73.12	74.28

TABLE II
ACCURACY (%) OF FOUR MTCNNS ON THE AVA DATASET.

δ	MTCNN #1	MTCNN #2	MTCNN #3	Enhanced MTCNN
0	76.15	75.91	75.92	76.58
1	75.90	75.81	75.37	76.04

with only aesthetic labels. It contains 20,278 images collected from *www.photo.net*. Each image is rated by at least 10 users to assess the aesthetic quality from one to seven. Due to some missing images in the dataset, we collect 17,232 images in all. From the overall images, 3000 images are randomly selected as the testing set, and the rest 15,232 images as the training set. For the ground truth labels, we follow [13] and choose the average score 5.0 as median aesthetic ratings. The images with an average score larger than $5+\delta$ are designated as high quality images, those with an average score smaller than $5-\delta$ as low quality images. We set δ to 0 in the experiment. Aesthetic quality assessment with $\delta = 0$ is more challenging than that with $\delta > 0$ [25].

B. Evaluating the Effectiveness of Keeping Balance Strategy

In the objective function, λ is used to control the contributions from semantic information. To validate our strategy of keeping the influence of two tasks balanced, we implement our MTCNN #1 with our strategy $\lambda = 1/M$ (here $\lambda = 1/29$) and we also compare the experimental results of MTCNN #1 with $\lambda = 0$, $\lambda = 2/29$, $\lambda = 1$ and early stopping strategy (shown in Table I). By comparing the results with or without the supervision of semantic labels, the MTCNN #1 with $\lambda \neq 0$ performs better than that with $\lambda = 0$. This indicates the supervision is effective. What's more, the results shown in Table I demonstrate that our strategy $\lambda = 1/29$ performs best on both values of δ. When $\lambda = 1/29$, the aesthetic and semantic tasks have same effect on the process of back propagation. Therefore the effectiveness of our strategy is verified.

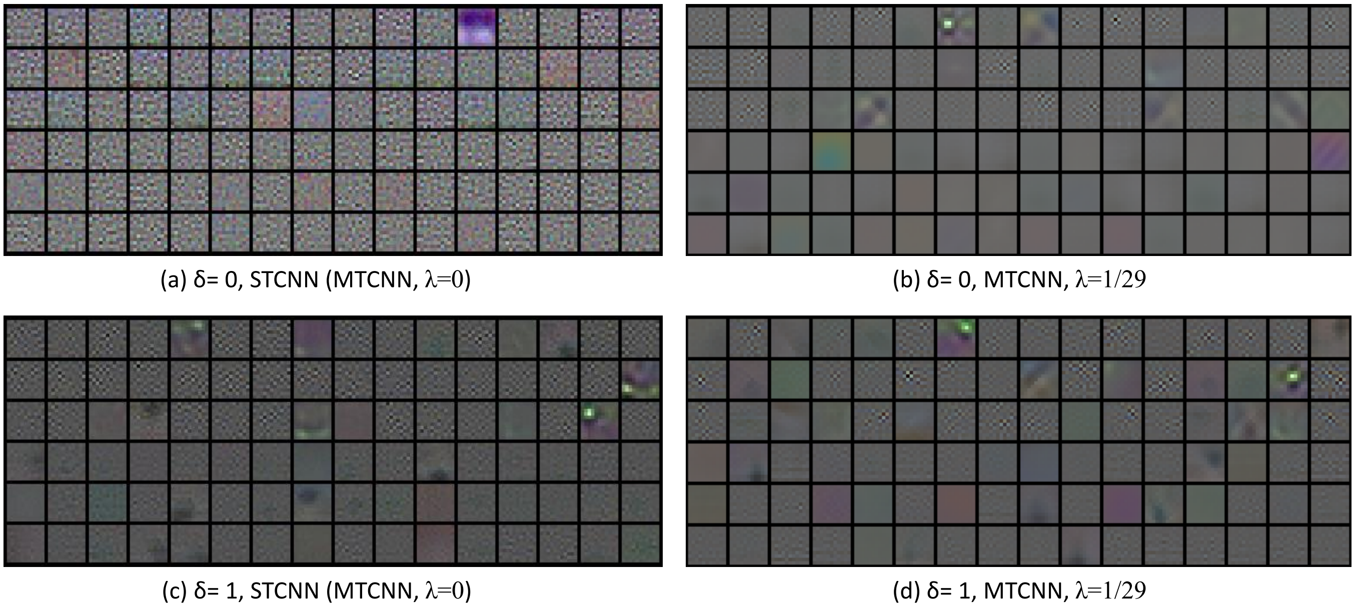
To further demonstrate the effectiveness of our MTCNN with our strategy, we also analyze the accuracy on each semantic tag using MTCNN #1 with different setting of λ in Fig. 5. As shown, our MTCNN #1 with $\lambda = 1/29$ performs best on overall images and most semantic tags. We also observe that different results are achieved on various semantic tags with the same method, and different improvements with MTCNNs are also different on various semantic tags. For example, the semantic tags "Family" and "Snapshot" obtain an great improvement with different methods.

C. Evaluating the impact of network architectures

To evaluate the impact of network architectures, we analyze the results with the three basic MTCNNs with $\lambda = 1/M$ and enhanced MTCNN with $\lambda = 2/M$ (shown in Table II). We can see that our enhanced MTCNN for the main task performs best. For the enhanced MTCNN, under the premise of ensuring the effect of semantic information in the whole network, we enhance the aesthetic supervision in the two bottom layers. Experimental results also show that MTCNN #1 performs best in the three basic MTCNNs. Comparing the MTCNN #1 and MTCNN #2, we can see that late splitting obtains better performance for aesthetic quality assessment and semantic information is helpful for aesthetic representation learning. This also demonstrates that the more supervision semantic labels makes on the aesthetic feature learning, the better performance our MTCNN achieves. It also reveals that the low-level features of MTCNN #3 can still perform well.

D. Evaluating the Benefits of Semantic Information

To evaluate our MTCNNs with the help of semantic information for aesthetic classification, we compare our results of four MTCNNs with those of our single task CNN (STCNN,



(a) δ= 0, STCNN (MTCNN, λ=0)

(b) δ= 0, MTCNN, λ=1/29

(c) δ= 1, STCNN (MTCNN, λ=0)

(d) δ= 1, MTCNN, λ=1/29

Fig. 7. Learned filters in the first convolutional layer with STCNN for aesthetic task only and MCTNN #1 for the two tasks with both $\delta = 0$ and $\delta = 1$.

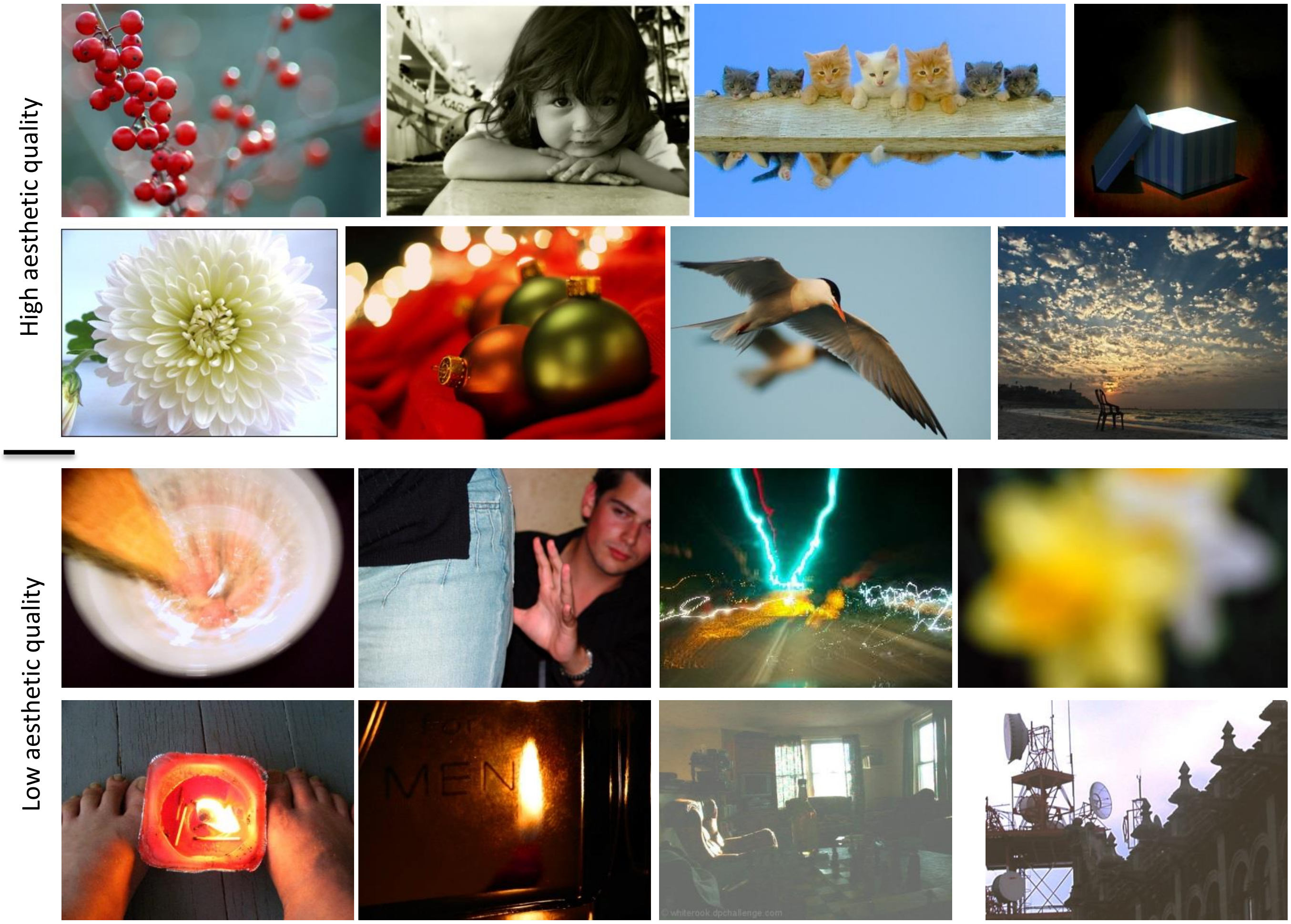


Fig. 8. Example test images correctly classified by MTCNN but incorrectly by STCNN in the AVA dataset. The labels of the images on the first and second rows are high aesthetic quality, and the labels of the images on the third and fourth rows are low aesthetic quality.

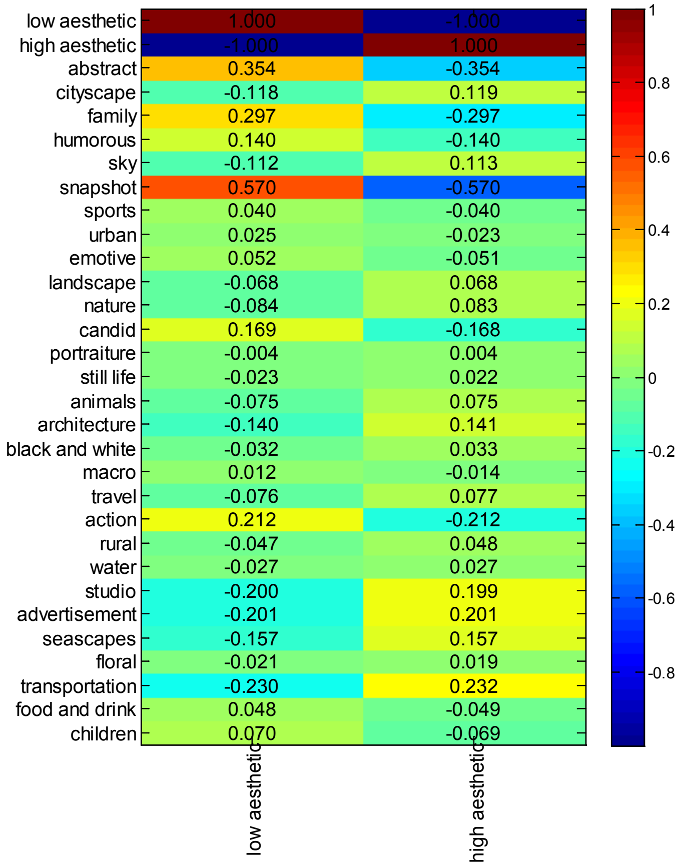


Fig. 9. Correlation in any two subtasks of aesthetic quality classification and semantic recognition learned by MTRLCNN #1 with $\delta = 0$.

MTCNN #1, $\lambda = 0$) on the AVA dataset with both values of δ. Shown in Table II and Table IV), all the four MTCNNs perform better than our STCNN especially when $\delta = 0$. Aesthetic quality classification with $\delta = 0$ is more challenging than that with $\delta = 1$ [25]. These results demonstrate the effectiveness of semantic information.

Furthermore, we also train a separate model for each semantic labels to assess aesthetic quality. Due to different number of images for different semantic labels, we only train four CNNs separately for “Landscape”, “Nature”, “Still Life” and ”Black and White”. The four labels have the most number of images in 29 labels. Here we call the CNNs trained separately for the four semantic labels “respective CNN”. For example, the respective CNN for “Landscape” is trained only with “Landscape” images for aesthetic categorization. Figure 6 shows the results with different methods for aesthetic classification on “Landscape”, “Nature”, “Still Life” and ”Black and White” separately with both value of δ. As shown in Fig. 6, all the MTCNNs outperform the respective CNN on each semantic labels, which also demonstrates the effectiveness of semantic information for representation learning. Moreover, MTCNNs don’t need to know the semantic labels of the testing images, while the respective CNNs have to know the semantic labels.

To qualitatively demonstrate the benefits of our MTCNN with semantic information, we show learned filters in the first convolutional layer with a STCNN for aesthetic task only and our MCTNN #1 with both $\delta = 0$ and $\delta = 1$ in Fig. 7. Compared to the filters learned without semantic information, the filters with semantic information are smoother, cleaner and more understandable. The proposed MTCNN can learn more color and high frequency edge information than STCNN. These differences can also be observed from the examples of test images correctly classified by MTCNN but misclassified by STCNN in Fig. 8. The high quality images often have more vivid color and clearer edge than low quality images. Most of the low quality images in Fig. 8 are blurred and dull. This indicates that the supervision of semantic labels for aesthetic feature learning is very beneficial, and aesthetic and semantic tasks are related to some extent.

TABLE III
ACCURACY (%) OF DIFFERENT NETWORK WITH OR WITHOUT RELATIONSHIP LEARNING ON THE AVA DATASET.

Architecture	MTCNN #1	AlexNet_FT	VGG Net_FT	ResNet_FT
MTCNN	76.15	76.70	77.73	**78.56**
MTRLCNN	76.56	77.35	78.46	**79.08**

To exploit the semantic information in the Imagenet, we select the late splitting multi-task network (such as MTCNN #1) and replace the MTCNN #1 architecture from Layer 1 to Layer 6 in Fig. 2 with AlexNet [27], VGG Net [51] or ResNet [52] respectively. That is because that the MTCNN #1 performs best in the three basic MTCNNs. The networks are initialized with models pretrained on Imagenet and finetuned with the training data labeled with aesthetic labels and semantic labels. Table III shows the results of the three MTCNN networks (AlexNet_FT, VGG Net_FT and ResNet_FT) with finetuning. It demonstrates the effectiveness of semantic information in Imagenet dataset. By comparing among three pre-trained networks, especially the ResNet [52], the deeper network learns more semantic representation and performs better for aesthetic quality assessment by transfer learning.

E. Inter Tasks Correlation Analysis

To further demonstrate the effectiveness of semantic information and investigate how semantic information influence aesthetic task again, we analyze the correlation between the two tasks. Since each column vector of task-specific matrix $W = [W_a, W_s]$ in the network corresponds to the parameters of a subtask, we use the learned covariance matrix Ω and calculate the correlation coefficient between any two subtasks [56]. Shown in layer 7 of Fig. 2 in our problem, the aesthetic classification task has two subtasks: high aesthetic and low aesthetic, the semantic recognition task has 29 subtasks. Figure 9 presents the correlation between the aesthetic subtasks and sematnic subtasks learned by MTRLCNN #1 with $\delta = 0$, which also verifies that semantic information is beneficial for aesthetic estimation. Seen from Fig. 9, a low aesthetic task has high negative correlation with a high aesthetic task. We can also see that the aesthetic tasks have high correlation with certain semantic attributes. For instance, the semantic tags “Snapshot” and “Candid” recognition has high positive correlation with the low aesthetic task. In real

TABLE IV
ACCURACY (%) OF DIFFERENT METHODS ON THE AVA DATASET.

δ	Our STCNN	MTCNN #1	MTRLCNN AlexNet_FT	MTRLCNN VGG Net_FT	MTRLCNN ResNet_FT	[25]	SCNN [29]	RDCNN [29]	DMA-Net [31]	MNA-CNN [32] (VGG Net_FT)
0	72.19	76.15	77.35	78.46	**79.08**	66.7	71.20	74.46	75.41	77.4
1	75.13	75.90	76.80	77.41	**77.71**	67.0	68.63	73.70	–	76.5

word, most of “Snapshot” and “Candid” images are usually regarded as low aesthetic quality images. While “Advertisement” and “Seascapes” recognition has positive correlation with the high aesthetic task. This accords with the knowledge that most of “Seascapes” and “Advertisement” images are usually taken as high aesthetic quality images. In addition, Fig. 9 can also visualize the correlation in different semantic tag recognitions. We also present the results of networks with or without relationship learning for aesthetic quality assessment in Table III, which validates the task relationship learning.

F. Comparison with Other State-of-the-art Methods

To further validate our method with semantic information for aesthetic classification, we compare our results with those of the state-of-the-art methods in [25], [29], [31], [32] on the AVA dataset. Shown in Table II and Table IV, all the multi-task models perform better than the method in [25], SCNN [29], and RDCNN [29] in on both values of δ. The method in [25] is the baseline of the AVA dataset and is implemented by extracting fisher vector (FV) descriptors [57] on the top of SIFT [16] information and SVM classifier [58]. SCNN is a single-column CNN, and RDCNN is a double-column CNN with an aesthetic column and a pretrained style column. Our results of MTRLCNN with VGG net and ResNet finetuning outperform the state-of-the-art method [32]. Thus, these results in Table II and Table IV illustrate the effectiveness of our method with semantic recognition task.

Since the name list of 20,000 testing images used in [25], [29], [31], [32] are unavailable, the 20,000 images for testing in this paper maybe potentially different from the 20,000 testing images in [25], [29], [31], [32]. Thus, we performed 4 times with similar operation (20,000 images are randomly selected for testing at each time) for MTCNN #1 ($\lambda = 1/29, \delta = 0$). The mean and variance (76.25%, 0.0066) are close to our 76.15%, which shows the robustness of our method. In addition, in this paper we selects 185,751 training images according to some rules, including the rule that all images need to have at least one semantic tag. It seems that the our training set is more clean than the 230,000 training images in [25], [29], [31], [32] and maybe helpful. To clarify how much benefit our method training with a “clean” set, we implement the baseline model (STCNN) trained on the full training set of 230,000 images. The accuracies on the same test set are 72.20% ($\delta = 0$), 75.27% ($\delta = 1$) and close to 72.19% ($\delta = 0$), 75.15% ($\delta = 1$) with a “clean” set. It seems that training with a “clean” set does not help the current method. This also demonstrates that our multi-task models with smaller training data can still outperform the state-of-the-art methods.

TABLE V
ACCURACY (%) OF DIFFERENT METHODS ON THE PHOTO.NET DATASET.

δ	GIST_SVM	FV_SIFT_SVM	STCNN	STCNN_FT	MTCNN #1_FT
0	59.90	60.80	61.00	62.10	**65.20**

Although our goal is to improve the performance of aesthetic quality assessment without considering the evaluation of semantic task, we also give the 64.89% Average Precision of MTCNN#1 ($\lambda = 1/29, \delta = 0$) and 67.44% of MTRLCNN with ResNet_FT ($\lambda = 1/29, \delta = 0$).

G. Evaluating the Transfer Learning for Photo.net Dataset

To utilize the semantic information for the dataset with only aesthetic labels, we transfer the learned representation with both aesthetic labels and semantic labels for the dataset with only aesthetic labels. In this paper, we exploit the learned representation with aesthetic and semantic labels from AVA dataset in MTCNN #1 and finetune it with Photo.net dataset with only aesthetic labels. We call this model as MTCNN #1_FT. To validate the effectiveness of transferred representation with semantic information, we finetune the pretrained STCNN model on AVA dataset with only aesthetic labels for Photo.net dataset (STCNN_FT). Moreover, we also train a STCNN on Photo.net dataset without finetuning. Furthermore, we implement the GIST descriptors [59] and FV on the top of SIFT with a SVM classifier (GIST_SVM and FV_SIFT_SVM). Table V shows the accuracy of these methods on Photo.net dataset. Fig. 10 visualizes some testing images correctly classified by MTCNN #1_FT but incorrectly by STCNN_FT in the Photo.net dataset. These reveal the effectiveness of transfer learning with semantic information.

V. CONCLUSION AND FUTURE WORK

In this paper, we have employed the semantic information to help discover representations for aesthetic quality assessment by formulating an end-to-end multi-task deep learning framework. Aesthetic quality assessment has not been taken as an isolation problem. To make full use of the semantic information and investigate how semantic information influence aesthetic task, four MTCNNs have been explored to learn the aesthetic representation jointly with the supervision of aesthetic and semantic labels. At the same time, a strategy of keeping the effect of two tasks balanced is presented to optimize the parameters of our multi-task networks. In addition, task relationship learning is modeled in the multi-task framework and the correlations in the two tasks have

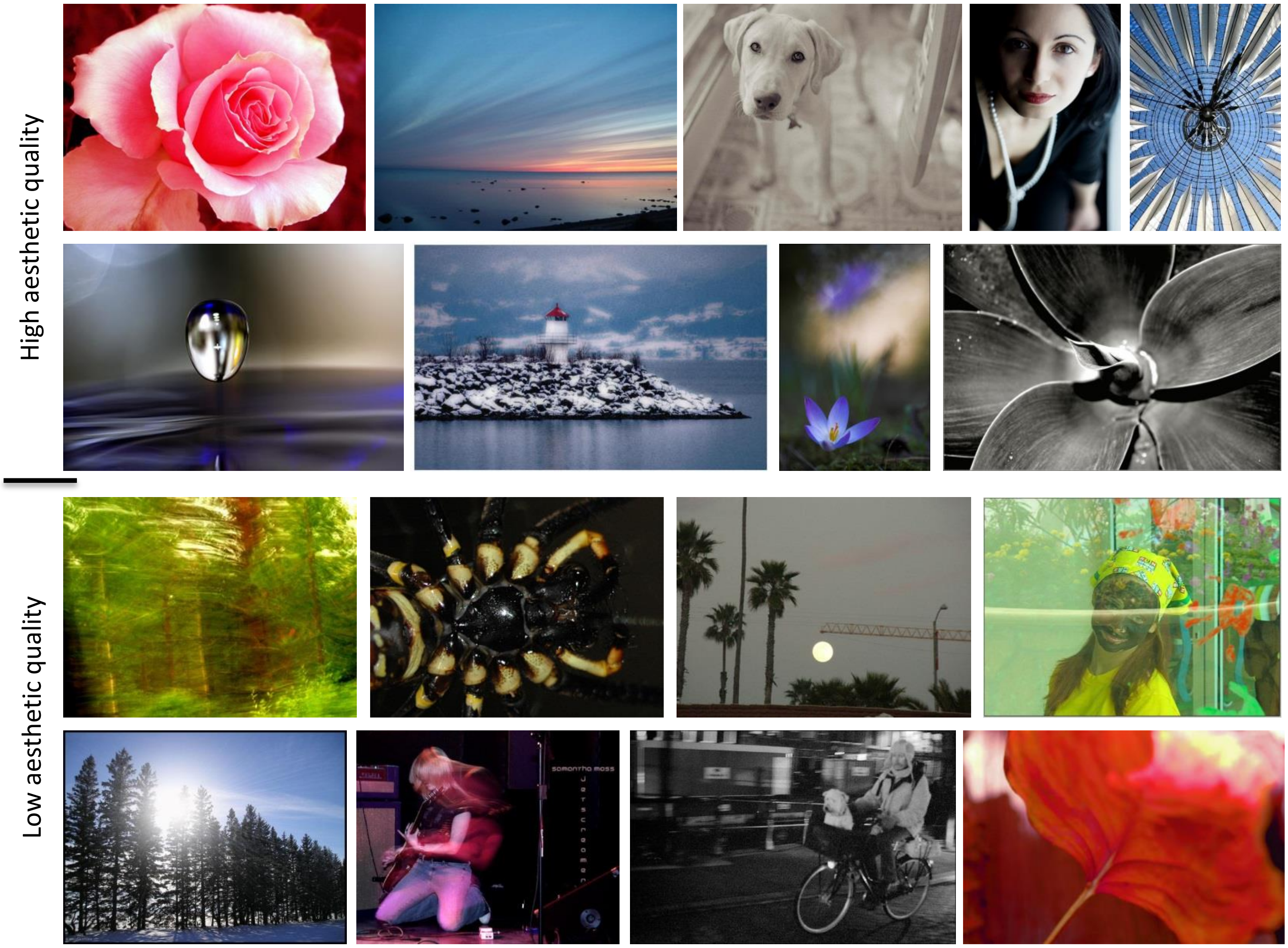


Fig. 10. Example test images correctly classified by MTCNN #1_FT but incorrectly by STCNN_FT in the Photo.net dataset. The labels of the images on the first and second rows are high aesthetic quality, and the labels of the images on the third and fourth rows are low aesthetic quality.

been learned to investigate the role of semantic recognition in aesthetic quality assessment. Experimental results have shown that our method performs better than the state-of-the-art methods. It is demonstrated that the semantic information is beneficial to aesthetic feature learning and the high-level features in the network play an important role in aesthetic quality assessment.

Although the proposed multi-task framework results in state-of-the-art results on the challenging dataset, how to perform aesthetic quality assessment like a human brain is still an ongoing issue. Future work is to explore other possible solutions to efficiently utilize the aesthetic and semantic information in a brain-like way. Another possible trend is to discover more possible and potential factors to affect aesthetic quality assessment.

References

[1] 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 Process.*, 2008, pp. 105–108.

[2] 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.

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

[4] L. Marchesotti, N. Murray, and F. Perronnin, "Discovering beautiful attributes for aesthetic image analysis," *Int. J. Comput. Vis.*, vol. 113, no. 3, pp. 246–266, Jul. 2015.

[5] E. Siahaan, A. Hanjalic, and J. Redi, "A reliable methodology to collect ground truth data of image aesthetic appeal," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1338–1350, July 2016.

[6] C. Segalin, A. Perina, M. Cristani, and A. Vinciarelli, "The pictures we like are our image: Continuous mapping of favorite pictures into self-assessed and attributed personality traits," *IEEE Trans. Affect. Comput.*, vol. PP, no. 99, pp. 1–1, 2016.

[7] J. Tarvainen, M. Sjöberg, S. Westman, J. Laaksonen, and P. Oittinen, "Content-based prediction of movie style, aesthetics, and affect: Data set and baseline experiments," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2085–2098, Dec. 2014.

[8] T.-S. Park and B.-T. Zhang, "Consensus analysis and modeling of visual aesthetic perception," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 272–285, Jul. 2015.

[9] R. Datta, J. Li, and J. Z. Wang, "Learning the consensus on visual quality for next-generation image management," in *Proc. ACM Int. Conf. Multimedia*, 2007, pp. 533–536.

[10] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 419–426.

[11] R. Hong, L. Zhang, and D. Tao, "Unified photo enhancement by discovering aesthetic communities from flickr," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1124–1135, Mar. 2016.

[12] L. Zhang, Y. Gao, R. Ji, Y. Xia, Q. Dai, and X. Li, "Actively learning human gaze shifting paths for semantics-aware photo cropping," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2235–2245, May 2014.

[13] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 288–301.

[14] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 386–399.

[15] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1657–1664.

[16] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1784–1791.

[17] H.-H. Yeh, C.-Y. Yang, M.-S. Lee, and C.-S. Chen, "Video aesthetic quality assessment by temporal integration of photo-and motion-based features," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1944–1957, Dec. 2013.

[18] Y. Wang, Q. Dai, R. Feng, and Y.-G. Jiang, "Beauty is here: Evaluating aesthetics in videos using multimodal features and free training data," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 369–372.

[19] 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, Mar. 2014.

[20] O. Wu, W. Hu, and J. Gao, "Learning to predict the perceived visual quality of photos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 225–232.

[21] L. Zhang, Y. Gao, C. Zhang, H. Zhang, Q. Tian, and R. Zimmermann, "Perception-guided multimodal feature fusion for photo aesthetics assessment," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 237–246.

[22] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 33–40.

[23] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22, 2004, pp. 1–2.

[24] T. S. Jaakkola, D. Haussler *et al.*, "Exploiting generative models in discriminative classifiers," *Proc. Adv. Neural Inf. Process. Syst.*, pp. 487–493, 1999.

[25] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2408–2415.

[26] L. Marchesotti, F. Perronnin, and F. Meylan, "Learning beautiful (and ugly) attributes." in *Proc. Brit. Mach. Vis. Conf.*, vol. 7, 2013, pp. 1–11.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[28] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[29] 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.

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

[31] 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. Comput. Vis.*, 2015, pp. 990–998.

[32] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 497–506.

[33] C. Mullin, G. Hayn-Leichsenring, and J. Wagemans, "There is beauty in gist: An investigation of aesthetic perception in rapidly presented scenes," *J. Vision*, vol. 15, no. 12, pp. 123–123, 2015.

[34] P. J. Locher, "The aesthetic experience with visual art at first glance," in *Investigations Into the Phenomenology and the Ontology of the Work of Art*, 2015, pp. 75–88.

[35] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[36] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 676 – 684.

[37] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 94–108.

[38] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2206–2213.

[39] Y. Niu and F. Liu, "What makes a professional video? a computational aesthetics approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 7, pp. 1037–1049, Jul. 2012.

[40] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3707–3715.

[41] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task cnn model for attribute prediction," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1949–1959, Nov. 2015.

[42] S. Zhang, R. He, Z. Sun, and T. Tan, "Multi-task convnet for blind face inpainting with application to face verification," in *Proc. International Conference on Biometrics*, 2016.

[43] W. Zhang, R. Li, T. Zeng, Q. Sun, S. Kumar, J. Ye, and S. Ji, "Deep model based transfer and multi-task learning for biological image analysis," in *Proc. KDD*, 2015, pp. 1475–1484.

[44] X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y.-Y. Wang, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," *Proc. NAACL*, 2015.

[45] Y. Zhang and D. Y. Yeung, "A convex formulation for learning task relationships in multi-task learning," in *Proc. Uncertain. Artif. Intell.*, 2010.

[46] A. Saha, P. Rai, S. Venkatasubramanian, and H. Daume, "Online learning of multiple tasks and their relationships," in *Proc. International Conference on Artificial Intelligence and Statistics*, 2011, pp. 643–651.

[47] E. V. Bonilla, K. M. Chai, and C. Williams, "Multi-task gaussian process prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 153–160.

[48] A. K. Gupta and D. K. Nagar, *Matrix variate distributions*. CRC Press, 1999, vol. 104.

[49] R. Girshick, "Fast r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015.

[50] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[54] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

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

[56] X. Fan, A. Felsovalyi, S. A. Sivo, and S. C. Keenan, "Sas for monte carlo studies," *SAS Institute, Cary*, pp. 87–89, 2002.

[57] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[58] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, Apr. 2011.

[59] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, May 2001.