Image Aesthetic Assessment Assisted by Attributes through Adversarial Learning

Bowen Pan,[1] **Shangfei Wang,**[*,1,2] **Qisheng Jiang**[2]
Key Lab of Computing and Communication Software of Anhui Province
[1]School of Computer Science and Technology,
[2]School of Data Science, University of Science and Technology of China, Hefei, Anhui, P.R.China, 230027
bowenpan@mail.ustc.edu.cn; sfwang@ustc.edu.cn; qishengj@mail.ustc.edu.cn

Abstract

The inherent connections among aesthetic attributes and aesthetics are crucial for image aesthetic assessment, but have not been thoroughly explored yet. In this paper, we propose a novel image aesthetic assessment assisted by attributes through both representation-level and label-level. The attributes are used as privileged information, which is only required during training. Specifically, we first propose a multitask deep convolutional rating network to learn the aesthetic score and attributes simultaneously. The attributes are explored to construct better feature representations for aesthetic assessment through multi-task learning. After that, we introduce a discriminator to distinguish the predicted attributes and aesthetics of the multi-task deep network from the ground truth label distribution embedded in the training data. The multi-task deep network wants to output aesthetic score and attributes as close to the ground truth labels as possible. Thus the deep network and the discriminator compete with each other. Through adversarial learning, the attributes are explored to enforce the distribution of the predicted attributes and aesthetics to converge to the ground truth label distribution. Experimental results on two benchmark databases demonstrate the superiority of the proposed method to state of the art work.

Introduction

Image aesthetic assessment has attracted increasing attention in recent years due to its wide application in personal photo album management, automatic photo editing, and image retrieval. A typical flow of image aesthetic assessment consists of feature extraction and decision phase. For feature extraction, both hand-crafted features and deep features have been explored. Early works often use aesthetic specific features to represent the photographic rules, such as lighting, contrast, and global image layout. Recent works have turned to learn deep features to represent image content. For decision phase, classifiers, such as naive Bayes classifier, support vector machine, and deep classifier, have been adopted to distinguish high-quality from low-quality photos in the form of binary classification, and regressors, like support vector regressor, have been used to predict aesthetic score.

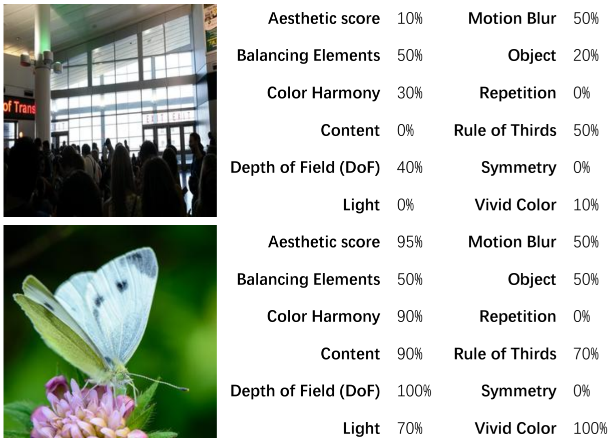


Figure 1: Two examples of aesthetic images (upper: low aesthetics; lower: high aesthetics) with respect to eleven assessment attributes. The ratings of the aesthetic score and attributes are written as percentage for convenience.

Although notable progresses have been achieved in assessing image aesthetics, it is still a big challenge to judge image aesthetics because of the subjectivity of beauty. To address it, several works try to leverage high level describable image attributes for image aesthetic assessment, since these attributes explicitly predict some of the possible image cues that a human might use to evaluate an image. Figure 1 displays two examples of aesthetic images with low and high aesthetic score, respectively. For the high aesthetic image, its nice content with good lighting and vivid color make it fascinating. While the low aesthetic image has the boring content with poor lighting and dull color. These probabilistic dependencies among aesthetics and attributes are crucial for image aesthetic assessment. Current work typically either uses attributes as low level features or middle level features, and thus fails to thoroughly leverage probabilistic dependencies among aesthetics and attributes for facilitating image aesthetic assessment.

Therefore, in this paper, we propose a novel attributes-enhanced image aesthetic assessment, where the attributes are used as privileged information (Vapnik and Vashist 2009). The attributes, which are only required during train-

[*]This is the corresponding author.

ing, are beneficial for aesthetic assessment from two perspectives. First, we propose a multi-task deep network to learn the aesthetic score and attributes simultaneously. Through multi-task learning, the attributes are explored to construct better feature representations. Second, we introduce a discriminator to distinct the predicted attributes and aesthetics from the ground truth ones. Through adversarial learning, the joint distribution inherent in the ground truth aesthetic scores and attributes is explored to further regularize the predicted attributes and aesthetics.

Related Works

A comprehensive survey on image aesthetic assessment can be found in (Joshi et al. 2011; Deng, Loy, and Tang 2017). In this section, we summarize several works that use attributes for image aesthetic assessment.

Early work tries to design features that might be related to how people judge aesthetic quality of photographs. For example, Ke *et al.* (Ke, Tang, and Jing 2006) summarized three distinguishing factors, i.e., simplicity, realism, and basic photographic technique, between high quality professional photos and low quality snapshots, and then designed spatial distribution of edges, color distribution, hue count, blur, contrast and brightness. Datta *et al.* (Datta et al. 2006) defined several quality specific features, including a low depth-of-field indicator, a colorfulness measure, a shape convexity score and a familiarity measure. Luo and Tang (Luo and Tang 2008) proposed several features based on the subject and background division. They first summarized several important criteria, i.e., composition, lighting, focus controlling, and color, which are used by professional photographers to improve photo quality through different treatment of the subject and the background. Then, they formulated clarity contrast feature, lighting feature, simplicity feature, composition geometry feature, and color harmony feature. Although their defined features are relevant to photography techniques, they can not totally capture high-level semantic attributes. This is an unavoidable weakness of engineering feature approaches.

Later work utilizes predicted attributes as middle-level representations. For example, Dhar *et al.* (Dhar, Ordonez, and Berg 2011) first predicted high level attributes from lower level features, and then predicted interestingness given high level attribute predictions. Kong *et al.* (Kong et al. 2016) proposed deep network to jointly learn photographic attributes and image content for photo aesthetics rating. The attribute predictor branch is fused with the aesthetic branch to produce a final attribute-adapted rating. Wang *et al.* (Wang et al. 2017) designed brain-inspired deep networks to learn attributes from features through the parallel supervised pathways, and then a high-level synthesis network is trained to transform those attributes into the overall aesthetics rating. Lu *et al.* (Lu et al. 2014) proposed to use attributes to regularize the feature learning and classier training for aesthetic quality categorization. Through employing attributes as middle-level representations, these works leverage attributes to learn features and train classifiers simultaneously. For middle-level representation approaches, attributes are usually manually annotated for the training data. During testing, the attributes are typically first predicted, and then the predicted attributes are used to measure aesthetics. Thus, the predicted errors of attributes can be propagated to the assessed aesthetics.

More recently, multi-task learning is adopted to train attributes and aesthetic score simultaneously. For example, Malu *et al.* (Malu, Bapi, and Indurkhya 2017) proposed to learn the aesthetic score and attributes jointly by using a deep convolution network with a merge-layer. The merge-layer collects pooled features of the convolution maps, and the aesthetic score and attributes are learned based on the merge-layer. Unlike middle-level representation approaches, a multi-task approach can avoid the predicted errors of attributes propagating to aesthetics. It can also exploit attributes for aesthetics assessment through the learned representations. However, it fails to model the distributions among attributes and aesthetics, which is crucial for image aesthetic assessment.

Therefore, in this paper, we propose a novel adversarial learning framework to model the joint distributions of aesthetic scores and attributes. Specifically, we use the aesthetics attributes as privileged information to train a deep convolutional rating network, which learns the aesthetic score and attributes simultaneously. Through multi-task learning, the attributes are beneficial for fine-tuning the feature representations. In order to further capture the correlation between the aesthetic score and attributes, a discriminator is introduced to distinguish the predictions from the ground truth and enforce the rating network to generate the predictions which are closer to the distribution of the ground truth.

Unlike engineering feature approaches, which fail to fully capture image content and attributes, the proposed method predicts attributes and aesthetic jointly from images using deep network. Therefore, the learned features can successfully represent image content due to the power of deep learning. Instead of using attributes as middle level representation, we predict attributes and aesthetic simultaneously. It avoids error propagation from attributes to aesthetic. Furthermore, we model the distribution among attributes to aesthetic through adversarial learning, and leverage such distribution to regularize aesthetic assessment.

Problem Statement

Let $\mathcal{D} = \left\{\left(\mathbf{x}^{(i)}, y^{(i)}, \mathbf{y}_a^{(i)}\right)\right\}_{i=1}^{N}$ denotes a training set of N training samples. Each training sample consists of a color image $\mathbf{x}^{(i)} \in \mathbb{R}^d$, an aesthetic score $y^{(i)} \in \mathbb{R}$ and K aesthetic attributes $\mathbf{y}_a^{(i)} \in \mathbb{R}^K$. Each aesthetic attribute can be either numerical ($y_k \in \mathbb{R}$) or binary ($y_k \in \{0, 1\}$). Due to the correlation between the aesthetic score and attributes, we want to build an attributes assisted model and further capture this correlation through both multi-task and adversarial learning. Given the training set $\mathcal{D}$, our goal is to learn a network $f : \mathbb{R}^d \to \mathbb{R}^{K+1}$, which outputs predictions of the aesthetic score and attributes simultaneously from color images. For convenience, we reformulate the dataset as $\mathcal{D} = \left\{\left(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right)\right\}_{i=1}^{N}$, where $\mathbf{y}^{(i)} \in \mathbb{R}^{K+1}$ is the concatenated vector of the aesthetic score and K attributes. This

$K+1$ dimensional vector space will be the space where we apply the adversarial learning as we will discuss later.

Proposed Method

The framework of the proposed method is summarized in Figure 2. As shown in Figure 2, there are two components in the proposed method: the rating network and the discriminator. The rating network outputs the prediction of the aesthetic score and attributes simultaneously and tries to use its prediction to fool the discriminator at the same time. The discriminator tries to distinguish the predictions from the real labels. During training, the aesthetic attributes are used as privileged information to construct better feature representations for aesthetic assessment. The correlation between the aesthetic score and attributes is learned through adversarial learning at the same time. During testing, the aesthetic score of an unknown image is predicted by the rating network. The predictions of the aesthetic attributes can be given if necessary.

Attributes Assisted Multi-task Rating Network

To take advantage of the assist of the aesthetic attributes, we use the aesthetic attributes as privileged information. Specifically, extra units are added upon the output layer to learn the aesthetic attributes, leading to a multi-task neural network. During back propagation, the gradient of the branch of aesthetic attributes will be used to adjust the feature representations below, which are shared with the aesthetic predictor. Therefore, the aesthetic attributes play the role of an assistant during training phase.

Formally, we build a multi-task deep convolutional rating network $\hat{\mathbf{y}} = R(\mathbf{x}; \theta_R)$ with $K+1$ output units corresponding to the aesthetic score and K aesthetic attributes. The objective function of the rating network is defined as follows:

$$J(\theta_R) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K+1} L\left(\hat{\mathbf{y}}_k^{(i)}, \mathbf{y}_k^{(i)}\right) \tag{1}$$

where the loss function L can be either squared error or binary cross entropy depending on the target variable is numerical or binary. The formulations of squared error and binary cross entropy are given in Eq. (2) and (3) respectively.

$$L_{\text{se}}(\hat{y}, y) = \frac{1}{2}(y - \hat{y})^2 \tag{2}$$

$$L_{\text{bce}}(\hat{y}, y) = -y \log \hat{y} - (1-y) \log(1-\hat{y}) \tag{3}$$

During training, both aesthetic score and attributes are learned with the supervised loss. The extra learning of the aesthetic attributes serves to regularize the learned representations of the rating network. During testing, the aesthetic attributes of the test images are not required.

Capturing Distributions of Attributes and Aesthetic for Aesthetic Assessment

Aesthetic score and attributes describe the same aspect of the color images and exhibit strong correlation. Although the rating network jointly predicts the aesthetic score and attributes through exploring shared representations, it does not explore label dependencies directly from ground truth labels. Therefore, there may exist a deviation between the distribution of the prediction and the ground truth label.

Inspired by the framework of Generative Adversarial Networks (GANs) (Goodfellow et al. 2014), we address this problem by modeling the correlation between the aesthetic score and attributes through adversarial learning. Specifically, we introduce a discriminator $y_D = D(\mathbf{y}; \theta_D)$ to distinguish the predictions from the real labels. For the rating network, we keep the supervised learning objective, i.e., minimize the error between the prediction and the ground truth label. In addition, we want the rating network to make predictions which fool the discriminator, leading to a new adversarial learning objective. Under these two learning objectives, the rating network will be expected to output reliable predictions which minimize the supervised loss and subject to real distribution of the aesthetic score and attributes at the same time.

Mathematically, the learning objective of the framework can be written as follows:

$$\begin{aligned}\min_{\theta_R} \max_{\theta_D} \frac{C}{N} \sum_{i=1}^{N} \left[\log D\left(\mathbf{y}^{(i)}\right) + \log\left(1 - D\left(R\left(\mathbf{x}^{(i)}\right)\right)\right)\right] \\ + \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K+1} L\left(R\left(\mathbf{x}^{(i)}\right)_k, \mathbf{y}_k^{(i)}\right)\end{aligned} \tag{4}$$

In the above learning objective, a weight coefficient C is added in order to control the proportion of the supervised objective and adversarial objective. Theoretically, if $C = 0$, the learning objective is equivalent to that of the multi-task network without the assist of discriminator. On the contrary, if $C \to +\infty$, the learning objective is the same as that of the original GAN.

Similar to the optimization procedure of GAN, the learning objective in Eq. (4) can not be optimized directly. The discriminator and the rating network are optimized alternately by fixing their opponents. According to the suggestion in (Goodfellow 2016), it is better for the rating network to minimize $-\log D(R(x))$ instead of minimizing $\log(1 - D(R(x)))$ in order to avoid the flat gradient. The learning objectives of the discriminator and the rating network are given in Eq. (5) and (6) respectively.

$$\max_{\theta_D} \frac{1}{N} \sum_{i=1}^{N} \left[\log D\left(\mathbf{y}^{(i)}\right) + \log\left(1 - D\left(R\left(\mathbf{x}^{(i)}\right)\right)\right)\right] \tag{5}$$

$$\begin{aligned}\min_{\theta_R} \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K+1} L\left(R\left(\mathbf{x}^{(i)}\right)_k, \mathbf{y}_k^{(i)}\right) \\ - \frac{C}{N} \sum_{i=1}^{N} \log D\left(R\left(\mathbf{x}^{(i)}\right)\right)\end{aligned} \tag{6}$$

Using the definition of binary cross entropy in Eq. (3), the learning objectives of the discriminator and the rating

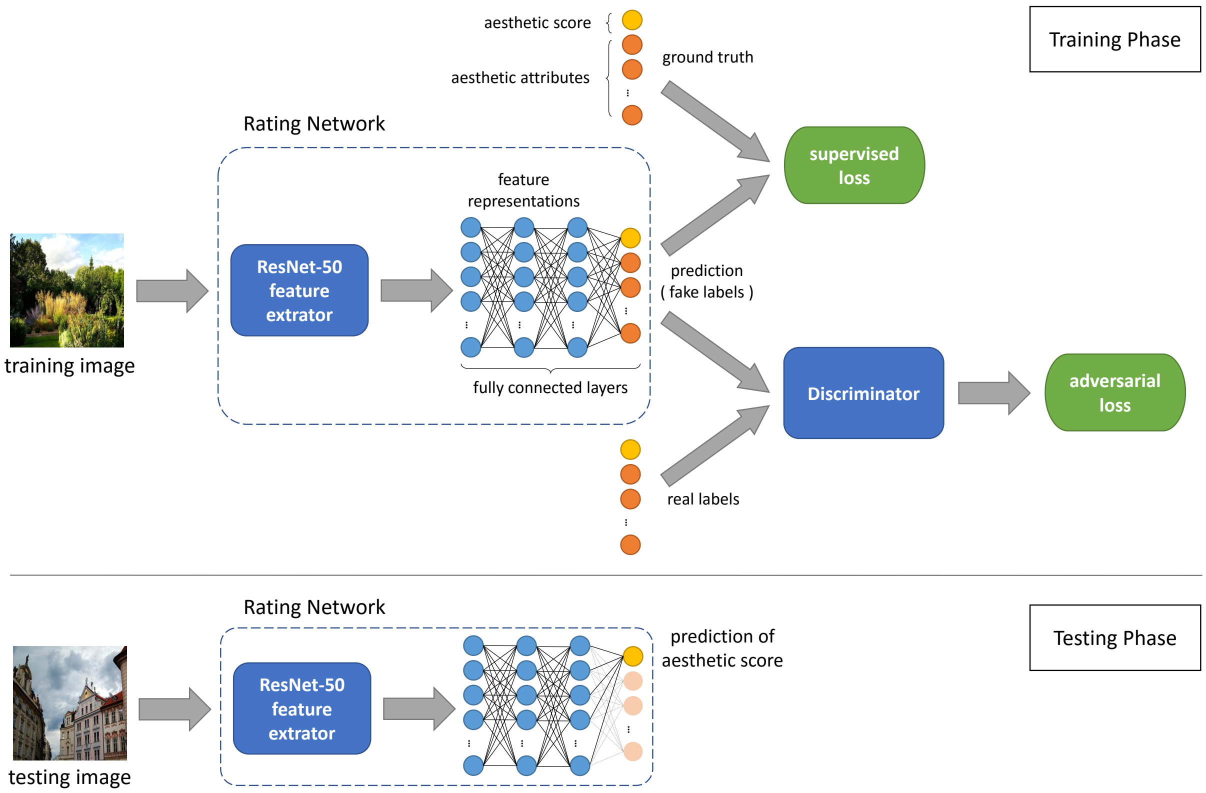


Figure 2: The framework of the proposed method.

network can be rewritten as follows:

$$J_D\left(\theta_D\right) = \frac{1}{N}\sum_{i=1}^{N}\left[L_{\text{bce}}\left(D\left(\mathbf{y}^{(i)}\right),1\right) + L_{\text{bce}}\left(D\left(R\left(\mathbf{x}^{(i)}\right)\right),0\right)\right] \tag{7}$$

$$J_R\left(\theta_R\right) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K+1} L\left(R\left(\mathbf{x}^{(i)}\right)_k, \mathbf{y}_k^{(i)}\right) + \frac{C}{N}\sum_{i=1}^{N} L_{\text{bce}}\left(D\left(R\left(\mathbf{x}^{(i)}\right)\right),1\right) \tag{8}$$

Optimization

The rating network plays the role of "generator" in the proposed framework. Therefore, we can apply the alternate optimization steps which are similar to the original GAN framework. Algorithm 1 outlines the learning procedure of the proposed method.

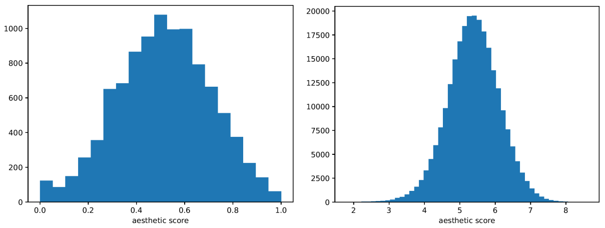
Experiment

Experimental Conditions

To the best of our knowledge, there are only two image aesthetic assessment databases containing aesthetic attributes: the Aesthetics and Attributes database (AADB) (Kong et al.

Algorithm 1 The learning algorithm of the attributes assisted image aesthetic assessment framework.

```
Require: The image aesthetic assessment dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N}$, the number of steps of updating discriminator $K_1$, the number of steps of updating rating network $K_2$, batch size $m$, learning rate of the discriminator $\eta_1$, learning rate of the rating network $\eta_2$.
Ensure: The rating network $R$.
Initialize parameters of the rating network $\theta_R$ and the discriminator $\theta_D$.
for number of training iterations do
    for $K_1$ steps do
        Sample a mini-batch of $m$ training images $\{\mathbf{x}^{(i)}\}_{i=1}^{m}$ from $\mathcal{D}$.
        Sample a mini-batch of $m$ real labels $\{\mathbf{y}^{(i)}\}_{i=1}^{m}$ from $\mathcal{D}$.
        Update the discriminator $D$ by gradient descent
        $\theta_D := \theta_D - \eta_1 \frac{\partial J_D(\theta_D)}{\partial \theta_D}$.
    end for
    for $K_2$ steps do
        Sample a mini-batch of $m$ training samples $\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{m}$ from $\mathcal{D}$.
        Update the rating network $R$ by gradient descent
        $\theta_R := \theta_R - \eta_2 \frac{\partial J_R(\theta_R)}{\partial \theta_R}$.
    end for
end for
```



(a) AADB

(b) AVA

Figure 3: The distributions of the aesthetic scores on the AADB and AVA databases.

2016) and the Aesthetics Visual Analysis database (AVA) (Murray, Marchesotti, and Perronnin 2012).

The AADB database contains a varied set of 10,000 photographic images downloaded from the Flickr website. Aesthetic quality score and eleven attributes are provided by five different individual raters using Amazon Mechanical Turk. The eleven attributes, including balancing elements, color harmony, content, depth of field (DoF), light, motion blur, object, repetition, rule of thirds, symmetry and vivid color, are selected according to the professional photographers' suggestions, which are closely related to image aesthetic judgements. The official partition for the AADB database are 8,500 images for training, 500 images for validation and 1,000 images for testing.

The AVA database contains about 250,000 images collected from the social network *www.dpchallenge.com*. Each image has about 200 aesthetic ratings ranging on a one-to-ten scale. A small portion of the images (about 14,000 images) are tagged with fourteen style attributes, i.e., complementary colors, duotones, HDR, image grain, light on white, long exposure, macro, motion blur, negative image, rule of thirds, shallow DoF, silhouettes, soft focus and vanishing point, and each style attribute is a binary variable. The official partition for the AVA database are 230,000 images for training, 20,000 images for testing.

The distributions of the aesthetic scores on two databases are shown in Figure 3. From Figure 3, we find that the distributions of the aesthetic scores are approximately Gaussian on both databases.

Given a color image, we first rescale the image so that the shorter side is of length 256. Then, a 224 × 224 patch is cropped randomly from the rescaled image on the training set for the purpose of data augmentation while the central 224 × 224 patch is cropped on the validation/test set. The aesthetic score and numerical aesthetic attributes are normalized to the interval of $[0, 1]$. Binary aesthetic attributes are converted to discrete values of 0 or 1. On the AADB database, the official train/validation/test split is adopted. On the AVA database, since there is only official train/test split, 20,000 images are selected randomly from the training set as the validation set so that the validation and test set have the same size. The ranking correlation measured by Spearman's ρ between the estimated aesthetic scores and the ground-truth scores is employed as performance metrics as in (Kong et al. 2016).

Table 1: Experimental results of image aesthetic assessment.

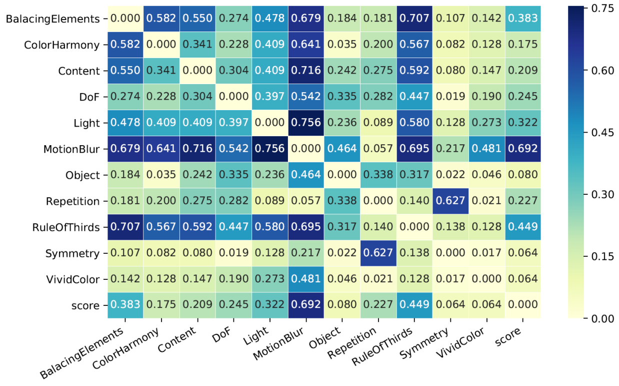
AADB database	
Methods	ρ
(Kong et al. 2016)	0.6782
(Hou, Yu, and Samaras 2017)	0.6889
(Malu, Bapi, and Indurkhya 2017)	0.689
Single-task Network	0.6833
Multi-task Network	0.6927
Ours	**0.7041**

AVA database	
Methods	ρ
(Kong et al. 2016)	0.5581
Single-task Network	0.6062
Multi-task Network	0.6187
Ours	**0.6313**

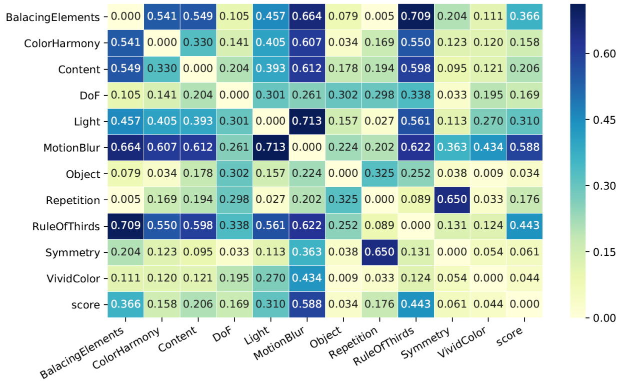
We conduct the experiments of image aesthetic assessment using three methods. The first one is the single-task network, which learns the mapping from the image to the aesthetic score. The second one is the multi-task network, which learns the mapping from the image to the aesthetic score and attributes simultaneously. The last one is our proposed method, in which multi-task learning for the aesthetic score and attributes is adopted and the correlation between them is learned in an adversarial manner.

We implement the proposed method and all of the related methods by PyTorch deep learning framework. Since all of the images in our experiment come from the natural scenes in daily life, it is beneficial to extract the feature representations by the deep models trained on the ImageNet database (Deng et al. 2009). Among all of the pre-trained models, the deep residual network (He et al. 2016) achieves the best performance due to its well-designed structure of residual blocks. For the rating network, we first extract feature representations from the pre-trained ResNet-50 and the size of the feature representations is 2048D. Upon the 2048D feature representations, we build two hidden full connected layers with ReLU activations. The sizes of these two layers are 512 and 128, respectively. The last layer is the output layer with sigmoid activation since all of the aesthetic score and aesthetic attributes are in $[0, 1]$. The size of the output layer is determined by the specific method and database. For the discriminator, a neural network with two hidden layers is used. The size of the hidden layer is eight.

We train the model using Adam algorithm (Kingma and Ba 2015) with a mini-batch size of 64. On the AVA database, a small portion of the images are tagged with aesthetic attributes. Therefore, we create each mini-batch containing images with and without aesthetic attributes for the proposed method and multi-task rating network. The loss of images with aesthetic attributes is computed by Eq. (1). The loss of images without aesthetic attributes is computed by the mean squared error of the aesthetic scores. The learning rate starts from 0.001 and is divided by 10 when the performance on the validation set plateaus.



(a) Multi-task network



(b) Ours

Figure 4: The absolute difference of correlation matrix on the AADB database.

Experimental Results and Analysis

Table 1 displays the experimental results of image aesthetic assessment on the AADB and AVA databases. As can be seen, our proposed method outperforms the single-task and multi-task networks. Specifically, the ranking Spearman coefficient achieved by our method is 0.0208 and 0.0114 higher those of the single-task and multi-task networks on the AADB database, respectively. On the AVA database, the ranking Spearman coefficient achieved by our method is 0.0251 and 0.0126 higher those of the single-task and multi-task networks, respectively. For the multi-task network, the aesthetic score and attributes are learned simultaneously and the additional task of learning aesthetic attributes assists the network in regularizing the learned representations below. The multi-task network achieves better performance than single-task network, indicating that using aesthetic attributes as privileged information is beneficial in image aesthetic assessment. However, the distribution of the predictions may be far away from that of the ground truth due to the lack of modeling the correlation between the aesthetic and attributes. For our proposed method, since we introduce a discriminator to distinguish the predictions from the ground truth, the rating network tries to fool the discriminator, which force the distributions of the prediction and ground truth to be closer. Therefore, our method achieves the best performance.

Evaluation of Adversarial Learning

In order to evaluate the effectiveness of adversarial learning, we analyse the joint distributions of the estimated aesthetic score and attributes versus the ground-truth quantitatively. We compute correlation matrix $M_{\text{true}} \in \mathbb{R}^{(K+1)\times(K+1)}$ of the ground truth of the samples in the test set on the AADB database. The correlation matrix M_{pred} of the prediction from a specific model is computed in the similar manner. Thus, the absolute difference between M_{true} and M_{pred} is a good measurement to evaluate the learning of the correlation between the aesthetic score and attributes.

Figure 4 displays the the absolute difference of correlation matrix with respect to the multi-task network and our proposed method, respectively. From Figure 4, we find that the average of the difference matrix in Figure 4a is 0.280 while the average of the difference matrix in Figure 4b is 0.250. It demonstrates that the distribution learned by our method is closer to the distribution of the ground truth. It is worth noting that it is easy to learning the correlation between aesthetic attributes like repetition, symmetry and vivid color, which describe the low-level aspect of the image. On the contrary, for those high-level aesthetic attributes like balancing elements, motion blur and rule of thirds, it is challenging to learn the correlation between them, resulting in large values in the difference matrix.

Analysis of Hyper Parameter

As can be seen in Eq. (8), the hyper parameter C controls the proportion of the supervised objective and adversarial objective. Theoretically, there exist an optimal value of C corresponding to the best trade-off between the supervised learning and adversarial learning. To explore the impact of hyper parameter C, we conduct experiments with different values of C. Take the AADB database for example, the experimental performance with respect to C is shown in Figure 5. As can be seen, when hyper parameter C gradually increases, the performance goes up at an early stage. After C is larger

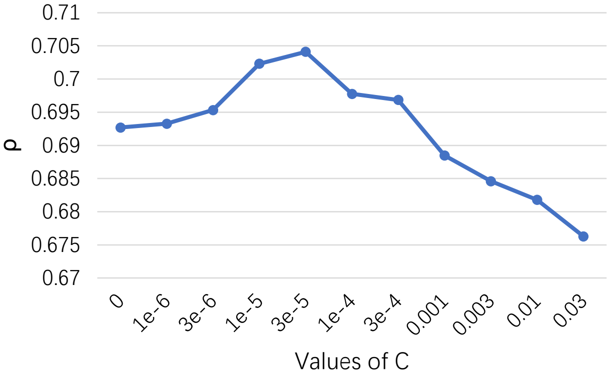
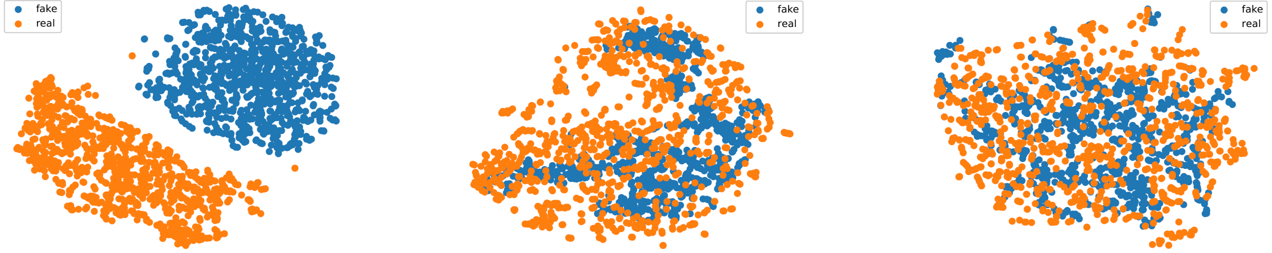


Figure 5: The experimental performance with respect to hyper parameter C on the AADB database.



(a) Multi-task network with random initialization

(b) Well-trained multi-task network

(c) Ours

Figure 6: Visualization of the output space on the AADB database.

than the optimal value, the performance degenerates quickly. This observation is consistent with the theoretical analysis.

Visualization of Adversarial Learning

To further evaluate the effect of the adversarial learning introduced in our model, we visualize the joint distributions of the estimated aesthetic score and attributes versus the ground-truth by t-SNE embedding (Maaten and Hinton 2008). The joint distributions of estimated aesthetic score and attributes which comes from three models are considered. The first one is the multi-task network with random initialization. The second one is the well-trained multi-task network. The last one is our proposed model. Only the visualization on the AADB database is plotted because the aesthetic score and attributes on the AADB database are all numerical.

Figure 6 displays all of the related visualization. The distribution of estimated aesthetic score and attributes is tagged “fake” in blue while the distribution of ground truth is tagged “real” in orange. From Figure 6a, we can obverse that the clusters of “real” and “fake” points are naturally separated when the multi-task network is initialized with random weights. When the multi-task network is trained with the loss function in Eq. (1), the cluster of “fake” points begins to overlap the cluster of “real” points as shown in Figure 6b. However, there still exist some areas where the “fake” points can not overlap. In Figure 6c, the “fake” points come from the rating network, which competes with the discriminator through adversarial learning. The areas of the overlap between the two clusters are larger compared with Figure 6c. Such visualization demonstrates that our proposed model can capture the correlation between the aesthetic score and attributes effectively. Thus the performance of the aesthetic score prediction can be further enhanced.

Comparison with related works

Three related works which achieve the state of the art performance on the benchmark databases, i.e., (Kong et al. 2016; Hou, Yu, and Samaras 2017; Malu, Bapi, and Indurkhya 2017), are compared with our method in Table 1. As can be seen, our method achieves the best performance on both AADB and AVA databases. Specifically, the ranking Spearman coefficient of our method is 0.0259, 0.0152 and 0.0151 higher than those of Kong *et al.*, Hou *et al.* and Malu *et al.*’s methods on the AADB database. On the AVA database, the ranking Spearman coefficient of our method is 0.0732 higher than that of Kong *et al.*’s method. In Kong *et al.*’s method, a branch is added to predict the aesthetic attributes upon the penultimate layer of the original network and the final aesthetic score is given based on the features of the aesthetic attributes and content. However, the correlation between the aesthetic score and attributes is totally ignored. While in our method the correlation is considered and captured by multi-task and adversarial learning. The joint learning of the aesthetic score and attributes ensures more robust feature representations and the adversarial learning closes the distributions between the predictions and ground truth labels. In (Hou, Yu, and Samaras 2017), Hou *et al.* proposed a new squared earth mover’s distance-based (EMD^2) loss which addresses the inter-class relationships. Although they demonstrated the superiority of the proposed EMD^2 loss, their network is intrinsically a single-task one without using the information of aesthetic attributes. We see that the multi-task network is able to outperform Hou *et al.*’s method, demonstrating the assistant role of the aesthetic attributes. In Malu *et al.*’s method, a merge-layer is introduced to collect pooled features of the convolution maps and the aesthetic score and attributes are learned based on the merge-layer. However, the network in Malu *et al.*’s method fails to capture the correlation between the aesthetic score and attributes. Our method addresses it through adversarial learning and achieves better performance than Malu *et al.*’s method.

Conclusions

In this paper, we propose an adversarial learning framework assisted by attributes for aesthetic assessment, which leverages aesthetic attributes as privileged information to construct a better predictor. A deep convolutional neural network with multiple outputs is adopted to learn the mapping from images to the aesthetic scores and attributes simultaneously during training phase. Adversarial learning is further introduced in order to capture the correlation between the aesthetic score and attributes, forcing the distributions of the prediction and the ground truth to be closer. Experimental results on the AADB and AVA databases demonstrate that our proposed method can capture the correlation between the aesthetic score and attributes effectively and then enhance the performance of the aesthetic assessment.

Acknowledgments

This work was supported in part by the Project from Anhui Science and Technology Agency under Grant 1508085SMF223, and the National Nature Science Foundation of China under Grant 61473270.

References

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. In Leonardis, A.; Bischof, H.; and Pinz, A., eds., *Computer Vision – ECCV 2006*, 288–301. Berlin, Heidelberg: Springer Berlin Heidelberg.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 248–255. IEEE.

Deng, Y.; Loy, C. C.; and Tang, X. 2017. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine* 34(4):80–106.

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

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Goodfellow, I. 2016. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hou, L.; Yu, C.-P.; and Samaras, D. 2017. Squared earth mover's distance-based loss for training deep neural networks. *NIPS Workshop*.

Joshi, D.; Datta, R.; Fedorovskaya, E.; Luong, Q.-T.; Wang, J. Z.; Li, J.; and Luo, J. 2011. Aesthetics and emotions in images. *IEEE Signal Processing Magazine* 28(5):94–115.

Ke, Y.; Tang, X.; and Jing, F. 2006. The design of high-level features for photo quality assessment. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, 419–426. IEEE.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. *ICLR*.

Kong, S.; Shen, X.; Lin, Z.; Mech, R.; and Fowlkes, C. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, 662–679. Springer.

Lu, X.; Lin, Z.; Jin, H.; Yang, J.; and Wang, J. Z. 2014. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, 457–466. ACM.

Luo, Y., and Tang, X. 2008. Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision*, 386–399. Springer.

Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(Nov):2579–2605.

Malu, G.; Bapi, R. S.; and Indurkhya, B. 2017. Learning photography aesthetics with deep cnns. *MAICS* 129–136.

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

Vapnik, V., and Vashist, A. 2009. A new learning paradigm: Learning using privileged information. *Neural networks* 22(5-6):544–557.

Wang, Z.; Liu, D.; Chang, S.; Dolcos, F.; Beck, D.; and Huang, T. 2017. Image aesthetics assessment using deep chatterjee's machine. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, 941–948. IEEE.