Semi-supervised Adversarial Learning for Attribute-Aware Photo Aesthetic Assessment

Yangyang Shu, Qian Li, Lingqiao Liu, and Guandong Xu

***Abstract*—Aesthetic attributes are crucial for aesthetics because they explicitly present some photo quality cues that a human expert might use to evaluate a photo's aesthetic quality. However, annotating aesthetic attributes is a time-consuming, costly, and error-prone task, which leads to the issue that photos available are partially annotated with attributes. To alleviate this issue, we propose a novel semi-supervised adversarial learning method for photo aesthetic assessment from partially attribute-annotated photos, which can greatly reduce the reliance on manual attribute annotation. Specifically, the proposed method consists of a score-attributes generator R, a photo generator G, and a discriminator D. The score-attributes generator learns the aesthetic score and attributes simultaneously to capture their dependencies and construct better feature representations. The photo generator reconstructs the photo by feeding aesthetic attributes, score, and informative feature representation. A discriminator is used to force the convergence of the features-attributes-score tuples generated from the score-attributes generator, the photo generator, and the ground-truth distribution in labeled data for training data. The proposed method significantly outperforms the state of the art, increasing the Spearman rank-order correlation coefficient (SRCC) from the existing best reported of 0.726 to 0.761 on *Aesthetic and attributes database* and 0.756 to 0.774 on *Aesthetic visual analysis database*, respectively.**

***Index Terms*—Semi-supervised adversarial learning, Aesthetic attributes, Photo aesthetic assessment**

I. Introduction

Photo Aesthetic Assessment (PAA) task aims to endow computers with the ability of perceiving aesthetics as human beings. The objective of this task is to automatically evaluate how beautiful the photo is from an aesthetic perspective. These assessments have many attractive applications, such as personal photo album management, automatic photo editing and photo retrieval. As a result, photo aesthetic assessment has become an increasingly popular research topic in recent years [1] [2] [3]. Although much progress has been achieved at learning a aesthetic assessment model, it is still a very challenging subjective task due to its heavy reliance on human perception of the photos.

To address this challenge, exploiting high-level descriptive aesthetic attributes has been an important channel for photo aesthetic assessment. This is mainly because human will not just give a high or low-quality judgment, or a numerical score but always describe many high-level descriptive aesthetic attributes such as color [4], content [5], light [6], motion blur [7], duotone [8], soft focus [9] of the photo. Figure 1 shows examples of how high-level attributes such as color, content, light can affect aesthetic qualities. The first and third rows include high-quality photos, whereas the second and fourth rows indicate low-quality photos. In this example, high-level aesthetic photos usually with harmonious color, abundant content, good light, vivid color, etc., make them fascinating. In the contrast, high-level attributes of disharmonious color, boring content, poor lighting, dull color, etc., determine the low aesthetic photos. Compared to aesthetic quality or score that describes global photo aesthetics, high-level descriptive aesthetic attributes represent subtle local photo aesthetics and thus are crucial for photo aesthetic assessment.

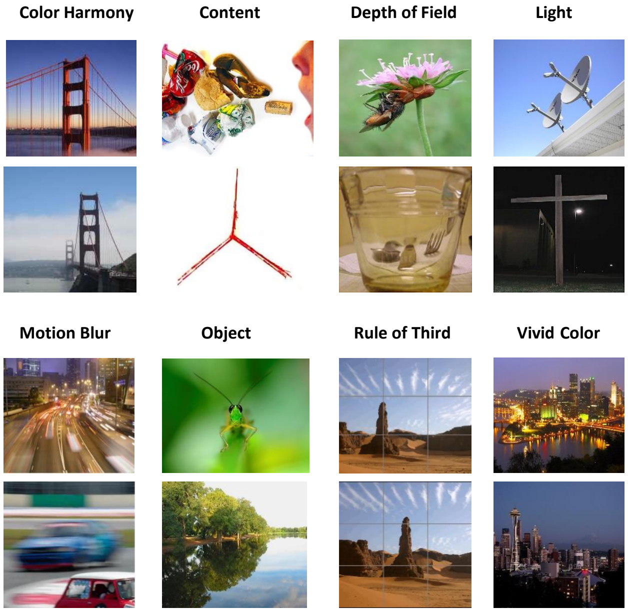

Fig. 1: Different aesthetic qualities w.r.t. different photo attributes. The first and third rows include photos with high aesthetic qualities. The second and fourth rows include photos with low aesthetic qualities.

Although fully attribute-annotated photos would be helpful for photo aesthetic assessment, collecting this data is time-consuming and error-prone. This results in incomplete and insufficient aesthetic attributes on photo databases. Recently, an increasing number of studies realizes the importance of high-level attributes for photo aesthetic assessment and focus

Y.Shu and G.Xu, are with Data Science and Machine Intelligence Lab and the Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW, 2007, Australia (e-mail: Yangyang.Shu@student.uts.edu.au, guandong.xu@uts.edu.au)

Q.Li is with School of Engineering, Computing and Mathematical Sciences, Curtin University, Perth, WA, 6102, Australia (e-mail: qli@curtin.edu.au)

L.Liu is with the School of Computer Science, University of Adelaide, Adelaide, SA 5005, Australia (e-mail: lingqiao.liu@adelaide.edu.au)

on exploiting the relations between attributes and aesthetics. Unfortunately, only a small number of existing attribute-annotated data is available due to the high labor cost. To reduce the reliance on manual attributes annotation, it is crucial to develop a method that only uses a small number of attribute-annotated data to leverage dependencies between high-level descriptive aesthetic attributes and aesthetics thoroughly.

Therefore, in this paper, we propose a semi-supervised adversarial learning for the attribute-aware photo aesthetic assessment method (SAGAN) to simultaneously learn the attribute and aesthetic score. The idea of the approach is constructing two generator networks and a discriminator network. Specifically, we first propose a score-attributes generator to learn the aesthetic score and attributes simultaneously. Second, we reconstruct photos by feeding the output of the score-attributes generator and minimize the reconstruction loss for all training data. Similarly, the aesthetic score and attributes are obtained by using the output of the photo generator. Third, we introduce a discriminator to align the joint distributions of predicted aesthetic attributes, predicted aesthetic score and the reconstructed photo features with the distribution of ground-truth. Through adversarial training, the joint inherent distribution in the ground truth is explored to further regularize the predicted attributes and aesthetics. Therefore, two generators leverage the partially available training data to predict the attribute-scores and aesthetic-scores, while the adversarial learning structure ensures the optimal network parameter training.

The contributions of this paper are as follows:

- We propose a semi-supervised adversarial learning method that can assess the photo aesthetics from partially attribute-annotated photos to reduce the reliance on aesthetic attributes annotation.
- We are the first to regard the generating processes of score-attributes and the photo as dual tasks [10], which generates informative feedback signals that benefit both tasks.
- We conduct extensive experiments to demonstrate the superiority of the proposed semi-supervised adversarial learning method compared to the state-of-the-art on two benchmark databases.

II. Related Work

In this Section, we focus on several works that utilize attributes for photo aesthetic assessment. We divide these works into three categories: low-level attributes, deep features, and high-level attributes. A comprehensive survey related to photo aesthetic assessment can be found in [11]–[13]. Furthermore, we discuss recent studies on semi-supervised adversarial learning.

A. Low-Level Attributes

In early researches, low-level attributes are extensively studied for aesthetic assessment. For example, Lowe [14] presents an engineering feature approach to transform image data into scale-invariant coordinates relative to local features. These features are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination. Ke *et al.* [4] propose three distinguishing factors, simplicity, realism, and basic photographic technique, making a photo high-quality or low-quality. Then, the spatial distribution of edges, color distribution, hue count, blur, contrast, and brightness are designed. Luo *et al.* [15] propose to use the subject region from a photo, then formulated clarity contrast feature, lighting feature, simplicity feature, and composition geometry feature based on the subject and background division. However, low-level features focus more on local hints instead of the global and contextual modeling of a photo. Therefore, they can not totally capture high-level attributes, which is the unavoidable weakness of engineering feature approaches.

B. Deep Features

With the development and application, deep features, extracted by deep convolutional neural networks, can capture the comprehensive understanding of the input photo. Lu *et al.* [16] use a novel double-column deep convolutional neural network to enable automatic feature learning. The proposed deep network unified the feature learning and classifier training, which helps to extract both global and local features of images. Kao *et al.* [17] find that semantic recognition task is vital for aesthetic quality prediction. Thus they cast the aesthetic prediction as the main task and captured the relations between semantic recognition task and aesthetic quality prediction via the proposed multi-task deep learning framework. Cui *et al.* [18] find that aesthetic perceiving is coupled with semantic understanding. Thus they develop a novel network framework, semantic-aware hybrid network (SANE), to harvest the information from object categorization and scene recognition via this network to improve the performance of aesthetic prediction. For their methods, attributes as input are needed in training. During testing, the attributes are still typically firstly predicted by the model or captured by database. This is usually complex and hardly satisfied in reality.

C. High-Level Attributes

1) Fully attribute-annotated photos: Several researchers have leveraged aesthetic attributes to facilitate photo aesthetic assessment. For example, Malu *et al.* [19] propose a novel multi-task deep convolution neural network (DCNN), which jointly learns high-level attributes along with the overall aesthetic score. In Malu's method, all aesthetic attributes are needed in their designed multi-task deep convolution neural network (DCNN). They develop a visualization technique to understand the internal representation of these attributes and qualitatively analyze the diverse and complex association with these different attributes. Shu *et al.* [20] proposes support vector regression with variational privileged information model and uses high-level descriptive attributes as privileged information to involve the learning process of their proposed method. In their method, all attributes are needed in the training phase, which provides the guidance of predicting the slack variable of support vector regression.

2) Partially attribute-annotated photos: Some studies focus on photo aesthetic from partially attribute-annotated samples. Pan *et al.* [21] propose a multi-task adversarial learning method to simultaneously learn aesthetic attributes and aesthetic score. Their method uses the available attributes as a target to help the network construct better feature representations for aesthetic assessment. Kong *et al.* [1] add a branch to predict the aesthetic attributes upon the penultimate layer of the original network. It is helpful only when attribute annotations are available. Then, the final aesthetic score is given based on the features of the aesthetic attributes and content. Although aesthetic attributes can be partially missing in these works, they still require a large amount of aesthetic attributes, and increasing the number of aesthetic attributes is definitely beneficial for photo aesthetic assessment.

D. Semi-Supervised Adversarial Learning

Recent years have seen a few works incorporating adversarial learning with semi-supervised learning. Lai *et al* [22] propose a generative adversarial network for learning optical flow in a semi-supervised manner where an adversarial loss was served as guidance for estimating optical flow from both labeled and unlabeled datasets to learn the structural patterns of the flow warp error. Miyato *et al.* [23] propose a virtual adversarial loss function without label information and applied this loss function into their proposed regularization method, which applies to semi-supervised learning. Wang *et al.* [24] propose a novel generative adversarial framework named CRGAN to solve the collaborative ranking problem. They use the pairwise rating comparison between different items as pairwise comparison setting. A discriminator is used to effectively enlarge the score difference between pairwise score comparison of items. The proposed method applied in partial pairwise preference annotation is considerable. Dong *et al.* [25] propose a margin generative adversarial network (MarginGAN) with three players, i.e. a generator, a discriminator and a classifier for semi-supervised learning. Their proposed network generate some pseudo labels, which are used for generated and unlabeled examples in training to make semi-supervised learning meet a variety of practical needs. All the above studies leverage adversarial learning for better input data or representations for semi-supervised learning, but ignore the informative feedback signals among different tasks. We are the first to exploit the feedback signals between the score-attributes generating task and the photo generating task. The proposed method explores the dependencies among photo features, aesthetic attributes and aesthetic score via adversarial learning and forces the joint distributions of the embedded set from the score-attributes generator and the photo generator to the ground-truth distribution.

III. Problem Statement

Denote a set of triples $T = \{x_i, a_i, y_i\}_{i=1}^{N}$ where feature vector $x_i \in \mathbb{R}^d$, aesthetic attributes $a_i \in \mathbb{R}^{d^*}$ and aesthetic score $y_i \in \mathbb{R}$. d and d^* are the dimensions of features and aesthetic attributes, respectively. N is the number of attribute-annotated training samples. Each aesthetic attribute can be either continuous ($a_{i,k} \in \mathbb{R}$) or binary ($a_{i,k} \in \{0,1\}$), k represents aesthetic attribute index. $C = \{x_j, y_j\}_{j=1}^{M}$ contains M training samples without attributes annotation. $S = T \cup C$ denotes $M+N$ dimensions training set containing all training samples. Let $A = \{x_i, y_i\}$ store all features and corresponding aesthetic scores in T, and $B = \{a_i, y_i\}$ store all aesthetic attribute vectors in T and their corresponding aesthetic scores. $V = A \cup C$ stores all training feature vectors and their corresponding aesthetic scores. Given the training set S, our goal is to jointly learn score-attributes generator R and photo generator G network where R network: $\mathbb{R}^d \to \mathbb{R}^{d^*+1}$ outputs the predicted aesthetic attributes and score simultaneously from feature vector, whereas G network: $\mathbb{R}^{d^R+d^*+1} \to \mathbb{R}^d$ outputs the generated feature vector from features representation x_R, attributes and score. d^R is the dimension of x_R from R network and formulated by "Decoding" network. Hence, we can exploit the relations between score-attributes generator R and photo generator G network to complement the missing attributes labels and improve the performance of two networks.

IV. Methodology

The framework of the proposed method is summarized in Figure 2. Specifically, through the score-attributes generator R, we get the predicted aesthetic attributes and aesthetic score $\{\hat{a}, \hat{y}\}$. Similarly, we sample data $\{a, y\}$ from B along with random noise $z \sim p_z(z)$ as the input of photo generator G, then we get the generated feature vector $\hat{x}^{'}$. The predicted aesthetic attributes and score $\{\hat{a}, \hat{y}\}$ along with the penultimate layer features x_R of score-attributes generator R are input into the photo generator G, and the output $\hat{x}$ is the reconstruction of x. Similarly, the generated feature $\hat{x}^{'}$ is input into score-attributes generator R, and the output of $\{\hat{a}^{'}, \hat{y}^{'}\}$ is the reconstruction of a and y. Decoding network is used to strengthen the penultimate layer features x_R holding the information of photo features and promote G to generate better feature representations. Thus, we consider three kinds of losses in the following subsections. Then, we give the learning objective functions and optimization.

A. Adversarial Loss

As shown in Figure 2, the discriminator D tries to distinguish the tuples $\{x, \hat{a}, \hat{y}\}$ and $\{\hat{x}^{'}, a, y\}$ from the ground truth distribution of $\{x, a, y\}$ embedded in the training data. $\{x, \hat{a}, \hat{y}\}$ and $\{\hat{x}^{'}, a, y\}$ are generated by score-attributes generator R and photo generator G, respectively to "fool" D, $\{x, a, y\}$ are sampled from set T. Under the competing process among score-attribute generator, photo generator and discriminator, the score-attributes generator R network will be expected to output predictions, which minimize the error. Specifically, the adversarial loss of discriminator is shown in Eq. 1.

$$\begin{aligned}\ell_{adv}^{D} = &-\mathbb{E}_{(x,a,y)\sim T}\log D(x,a,y)\\ &-\alpha\mathbb{E}_{(x,y)\sim V}\log(1-D(x,\hat{a},\hat{y}))\\ &-(1-\alpha)\mathbb{E}_{(a,y)\sim B, z\sim p_z(z)}\log(1-D(\hat{x}^{'},a,y)).\end{aligned} \tag{1}$$

where $\alpha \in [0,1]$ is a trade-off between the distribution of the pseudo-tuples generated from R and G network.

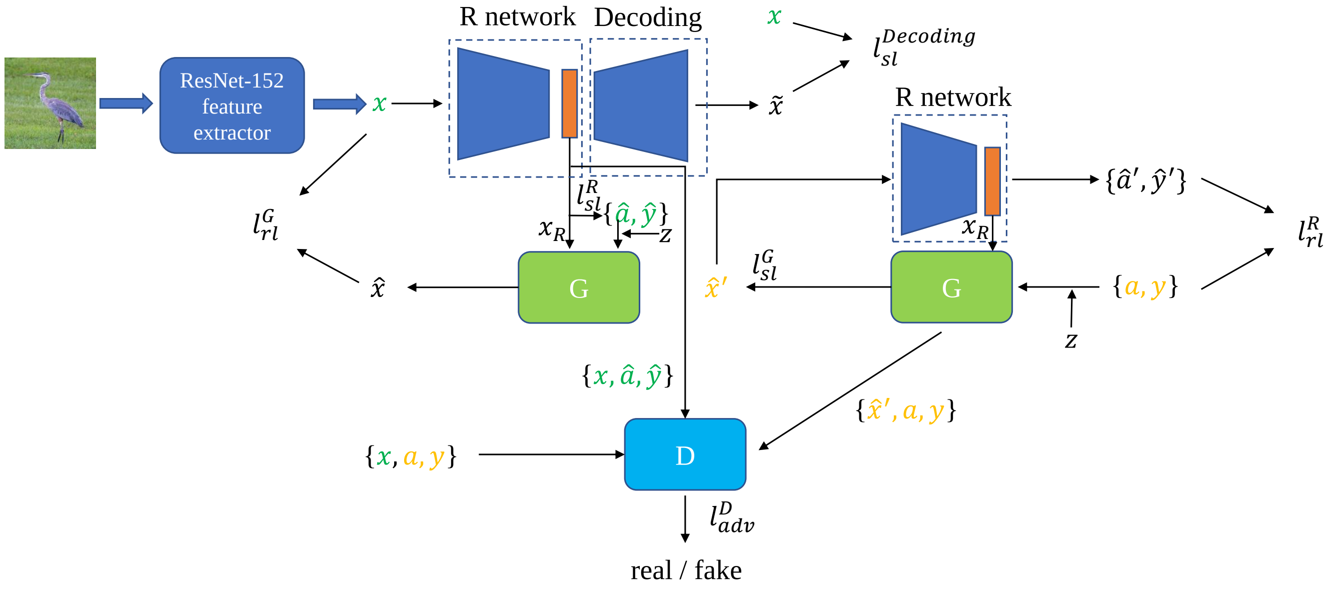

Fig. 2: The framework of the proposed semi-supervised adversarial learning.

B. Supervised Loss

Supervised loss contributes to the capacity of the network to predict the commonly occurring range. As shown in Figure 2, we use the supervised loss in the generated attributes and score $\{\hat{a}, \hat{y}\}$ in R network. Supervised losses are also used in the generated features $\tilde{x}$ and $\hat{x}^{'}$ in decoding network and G network respectively. We define them as:

$$\ell_{sl}^{R} = \mathbb{E}_{(x,a,y)\sim T} L(\{a, y\}, \{\hat{a}, \hat{y}\}), \tag{2}$$

$$\ell_{sl}^{Decoding} = \mathbb{E}_{(x,y)\sim V} L(x, \tilde{x}), \tag{3}$$

$$\ell_{sl}^{G} = \mathbb{E}_{(a,y)\sim B} L(x, \hat{x}^{'}), \tag{4}$$

where the loss function can be either squared error or binary cross entropy relying on the target variable is continuous or binary. The formulations of squared error and binary cross entropy are $L(y, \hat{y}) = (y - \hat{y})^2$ and $L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$, respectively.

C. Reconstruction Loss

We introduce the reconstruction loss and hope that the G network and R network can reconstruct photo information, aesthetic attributes and aesthetic score. The reconstruction losses for network G and R are shown in the following:

$$\ell_{rl}^{G} = \mathbb{E}_{(x,y)\sim V, z\sim p_z(z)} L(x, \hat{x}), \tag{5}$$

$$\ell_{rl}^{R} = \mathbb{E}_{(a,y)\sim B} L(\{a, y\}, \{\hat{a}^{'}, \hat{y}^{'}\}), \tag{6}$$

where the loss function L is squared error or binary cross entropy depending on the continuous or binary type of target variable.

D. The Learning Objectives of Each Network

We give the objective functions of D, R and G networks, respectively.

1) Gradient descent on D network:

$$\begin{aligned} \min\ _{\theta_D} [& -\mathbb{E}_{(x,a,y)\sim T} \log D(x, a, y) \\ & - \alpha \mathbb{E}_{(x,y)\sim V} \log(1 - D(x, \hat{a}, \hat{y})) \\ & - (1 - \alpha) \mathbb{E}_{(a,y)\sim B, z\sim p_z(z)} \log(1 - D(\hat{x}^{'}, a, y))]. \end{aligned} \tag{7}$$

2) Gradient descent on R network:

$$\begin{aligned} \min\ _{\theta_R} [& -\mathbb{E}_{(x,y)\sim V} \log(D(x, \hat{a}, \hat{y})) \\ & + \lambda_{sr} \mathbb{E}_{(x,a,y)\sim T} L(\{a, y\}, \{\hat{a}, \hat{y}\}) \\ & + \lambda_{rr} \mathbb{E}_{(a,y)\sim B} L(\{a, y\}, \{\hat{a}^{'}, \hat{y}^{'}\})] \end{aligned} \tag{8}$$

where λ_{sr} and λ_{rr} are weights of supervised loss and reconstruction loss respectively, in order to balance the tradeoff among losses.

3) Gradient descent on G network:

$$\begin{aligned} \min\ _{\theta_G} [& -\mathbb{E}_{(a,y)\sim B, z\sim p_z(z)} \log(D(\hat{x}^{'}, a, y)) \\ & + \lambda_{sg} \mathbb{E}_{(a,y)\sim B} L(x, \hat{x}^{'}) \\ & + \lambda_{rg} \mathbb{E}_{(x,y)\sim V, z\sim p_z(z)} L(x, \hat{x})] \end{aligned} \tag{9}$$

where λ_{sg} and λ_{rg} are weights of supervised loss and reconstruction loss respectively.

The optimization of the proposed objective function is similar to the original GAN framework [26]. We outline the proposed algorithm in Algorithm 1.

V. Experiments

To evaluate the effectiveness of the proposed method, we compare our method with the state-of-the-art methods on two benchmark databases of photo aesthetic assessment.

Algorithm 1 The training of semi-supervised adversarial learning of attributes in aesthetic assessment

```
Require: training sample $T$, $V$ and $B$, hyper parameters $\alpha$, $\lambda_{sr}$, $\lambda_{rr}$, $\lambda_{sg}$, and $\lambda_{rg}$, batch size $c$; the number of steps of updating
    discriminator network $K_1$, the number of steps of updating generator networks $K_2$, the number of training iterations $K$.
Ensure: score-attributes generator $R$ and photo generator $G$.
    Randomly initialize model parameters $\theta_D$, $\theta_R$ and $\theta_G$ respectively;
    for $K$ iterations do
      for $K_1$ steps do
        Sample a mini-batch of $c$ training data $\{x_i, a_i, y_i\}_{i=1}^{c}$ from $T$, sample a mini-batch of $c$ training data $\{x_i, y_i\}_{i=1}^{c}$ from
        $V$, sample a mini-batch of $c$ training data $\{a_i, y_i\}_{i=1}^{c}$ from $B$.
        Update discriminator $D$ by gradient descent:
          $\nabla_{\theta_D}[-\frac{1}{c}\sum_{i=1}^{c}((\log D(x,a,y) + \alpha\log(1 - D(x,\hat{a},\hat{y})) + (1-\alpha)\log(1 - D(\hat{x}^{'},a,y)))]$
      end for
      for $K_2$ steps do
        Sample a mini-batch of $c$ training data $\{x_i, y_i\}_{i=1}^{c}$ from $V$, $c_1$ ($c_1 \leq c$) samples of $\{x_i, y_i\}_{i=1}^{c}$ originally annotated
        with aesthetic attributes are used to construct $\{x_i, a_i, y_i\}_{i=1}^{c_1}$.
        Update score-attribute generator $R$ by gradient descent:
          $\nabla_{\theta_R}[-\frac{1}{c}\sum_{i=1}^{c}\log(D(x,\hat{a},\hat{y})) + \frac{\lambda_{sr}}{c_1}\sum_{i=1}^{c_1}L(\{a,y\},\{\hat{a},\hat{y}\}) + \frac{\lambda_{rr}}{c_1}\sum_{i=1}^{c_1}L(\{a,y\},\{\hat{a}^{'},\hat{y}^{'}\})]$
        Sample a mini-batch of $c$ training data $\{a_i, y_i\}_{i=1}^{c}$ from $B$, sample a mini-batch of $c$ training data $\{x_i, y_i\}_{i=1}^{c}$ from $V$.
        Update photo generator $G$ by gradient descent:
          $\nabla_{\theta_G}[-\frac{1}{c}\sum_{i=1}^{c}\log(D(\hat{x}^{'},a,y)) + \frac{\lambda_{sg}}{c}\sum_{i=1}^{c}L(x,\hat{x}^{'}) + \frac{\lambda_{rg}}{c}\sum_{i=1}^{c}L(x,\hat{x})]$
      end for
    end for
```

TABLE I: Attribute annotations on the AADB and AVA database.

Attribute (AADB)	Number	Attribute (AVA)	Number
Balancing elements	10,000	Complementary	949
Color Harmony	10,000	Duotones	1,301
Content	10,000	HDR	396
Depth of Field	10,000	Photo Grain	840
Light	10,000	Light on White	1,199
Motion Blur	10,000	Long Exposure	845
Object	10,000	Macro	1,698
Repetition	10,000	Motion Blur	609
Rule of Thirds	10,000	Negative Blur	959
Symmetry	10,000	Rule of Thirds	1,031
Vivid Color	10,000	Shallow DoF	710
-	-	Silhouettes	1,389
-	-	Soft Focus	1,479
-	-	Vanishing Point	674

A. Experimental Conditions

1) Datasets: In our experiments, we use two publicly available photo aesthetics databases: Aesthetics and Attributes database (AADB) [1] and Aesthetics Visual Analysis database (AVA) [27]. AADB database contains 10,000 photographic images from the Flickr website [1]. Aesthetic quality scores are continuous values in the interval of [0, 1]. The AADB database provides 11 attributes that are continuous values including balancing element, content, color harmony, depth of field, lighting, motion blur, object emphasis, rule of third, vivid color, repetition and symmetry. The official split of the AADB database is 8,500 images for training, 500 images for validation and the remaining 1,000 images for testing. The AVA database contains aesthetic quality scores in the interval [1, 10] and 14 style attributes in a small portion of photos (14,000/250,000), i.e., complementary, duotones, HDR, photo grain, light on white, long exposure, macro, motion blur, negative photo, rule of third shallow DOF, silhouettes, soft focus and vanishing point. These attributes are binary with discrete values of 0 or 1. Moreover, the AVA database is split into training set (230,000 samples) and testing set (20,000 samples). The official split of the AVA database does not include a validation set, and thus we have to construct the validation set by ourselves. We randomly sample 20,000 photos from the training set as the validation set, and the remaining 210,000 photos for training. The detailed attributes annotations on the AADB and AVA databases are listed in Table I.

2) Implementation Details: We implement our method using the PyTorch framework. In our experiments, photos are resized to 224×224 on both databases. We normalize the aesthetic score and continuous attributes to the interval [0, 1]. ResNet-152 [28] is used as the architecture to extract the feature representations. The pre-trained weights of ResNet-152 on Imagenet are used for initialization. Then 2048D feature representations are extracted from this pre-trained architecture. Score-attributes generator R, photo generator G, discriminator D and decoding network are parameterized by four-layer feed forward networks. The size of two hidden layers in score-attributes generator R are 512, 128 respectively. The last layer is the output layer with sigmoid activation and the size of the last layer is determined by the specific method and database, which are 12 and 15 on the AADB and AVA databases, respectively. For photo generator G and decoding network, the size of two hidden layer are both 128 and 512. The last layer of theses two networks is the the 2048D feature representations. For the discriminator, we also use the two hidden layers in a neural network with the size of 512, 128, respectively. Adam [29] optimizer with a mini-batch size of 64 is used to optimize R, G, D and decoding network in our experiments. Other hyper parameters, such as learning rate, training step K_1, K_2, and weight coefficients λ_{sr}, λ_{rr}, λ_{sg}, λ_{rg} are determined by a validation set.

3) Evaluation Metrics: The Spearman Rank-order Correlation Coefficient (SRCC) [1], Pearson Linear Correlation Coefficient (PLCC) [3] and accuracy [1] are used as metric evaluation. For all metrics, the larger the better.

4) Baselines: Firstly, we compare our method to attribute-aware methods. We choose the following eight popular methods for comparison. For a fair comparison, we re-implement RA-DCNN method by changing AlexNet [1] with ResNet framework. Among these methods, DCNN is supervised methods and the others are semi-supervised methods.

- Ranking and attribute deep convolutional rating network(RA-DCNN) includes additional activation layers in their proposed ranking network to encode informative attributes. Their methods regard prediction task as a source of side-information to regularize the weights of the network [1].
- Attribute-aware deep convolution neural network (DCNN) [19] jointly learns aesthetic attributes and aesthetic score simultaneously to learn the inherent representation of these attribute, and qualitatively analyze the diverse and complex association with these different attributes for aesthetic assessment.
- Adversarial deep convolutional rating network (ATTGAN) [21] proposes using deep rating network as a generator outputting attributes and score simultaneously. ATTGAN uses a discriminator to distinguish the generative attributes, score and ground truth by regularizing the distributions of high-level attributes and score.
- multi-task deep learning (MTDL) [30] proposes to use the personality features that are learned to modulate the aesthetic attributes. These attributes are used to predict the optimal generic photo aesthetics scores.
- Privileged information deep convolutional network (PI-DCNN) [31] uses attribute loss function, which explores the domain knowledge of the photo-based and photography-based attributes to improve the performance of aesthetic assessment.
- Support vector regression with variational privileged information(v-SVR+) [20] integrates aesthetic attributes as privileged information to predict the slack variable of support vector regression further regularizing the learning process and improving the learning performance of aesthetic prediction.
- Deep chatterjee's machine (DCM) [32] first learns aesthetic attributes via a parallel supervised pathway. Then they associate and transform those attributes into the overall aesthetic quality assessment.
- A multi-task framework (MTRLCNN) [17] is proposed to capture the correlations between semantic recognition and aesthetic quality prediction via incorporating the inter-task relationship learning.

Secondly, we compare our method to attribute-unaware methods. We choose the following seven popular methods for comparison.

- Squared earth mover's distance (Square-EMD) [33] is used to define loss functions for convolutional neural network training. They transform aesthetic database into a classification database, and leverage these relations between different classes via the squared earth mover's distance.
- A unified probabilistic formulation (AUPF) [34] uses an effective loss function to capture the inherent relation among binary classification, average score regression and score distribution prediction.
- Composition-aware network(CAN) [35] leverages the image composition information as the mutual dependency among its local regions to boost the performance of aesthetics.
- Convolutional ordinal regression forest (CORF) [36] obtains precise and stable global ordinal relationships via integrating ordinal regression and differentiable decision trees with a CNN for photo aesthetic assessment.
- Neural image assessment (NIMA) model [37] uses convolutional neural network to predict the distribution of human opinion scores which capture the high correlations between photo score and human perception.
- Gated peripheral-foveal convolutional neural networks (GPF-CNN) [38] encodes the holistic information to provide the attended regions, then extracts fine-grained features on these key regions.
- Multi-level spatially pooled features (MLSP) [3] are extracted from convolutional blocks to efficiently support full resolution images on variable input sizes.

5) Experiment Design: The number of attributes-annotated photos on the AADB and AVA database is 10,000 and 14,000, respectively. We randomly exclude attributes-annotated photos with four ratios, i.e., 10%, 20%, 30% and 40%. The proposed method employs aesthetic attributes as privileged information [39] to improve the learning performance of aesthetic assessment tasks. We compare the proposed method to a method that does not output aesthetic attributes in R network, referred to as SGAN. Besides, we compare the proposed method with another method that removes G network in our method, referred to as SAGAN-.

TABLE II: Within database experimental results of semi-supervised learning for aesthetic assessment with four missing rates of attribute-annotated photos on the AADB and AVA databases.

Database	Methods	10%			20%			30%			40%		
		SRCC	Accuracy	PLCC	SRCC	Accuracy	PLCC	SRCC	Accuracy	PLCC	SRCC	Accuracy	PLCC
AADB	v-SVR+	0.684	79.62%	0.676	0.662	77.55%	0.667	0.638	75.43%	0.640	0.620	74.09%	0.629
	RA-DCNN	0.665	77.54%	0.658	0.643	75.88%	0.644	0.621	74.64%	0.627	0.603	72.18%	0.597
	PI-DCNN	0.689	79.96%	0.686	0.674	78.02%	0.672	0.633	76.25%	0.628	0.602	72.68%	0.616
	ATTGAN	0.687	78.42%	0.692	0.674	77.65%	0.678	0.646	74.83%	0.652	0.605	72.11%	0.611
	SGAN	0.686	78.88%	0.683	0.664	76.52%	0.674	0.643	74.40%	0.655	0.635	73.69%	0.639
	SAGAN-	0.692	79.52%	0.702	0.677	78.13%	0.682	0.653	76.04%	0.660	0.648	75.52%	0.647
	SAGAN	**0.747**	**81.47**%	**0.752**	**0.722**	**80.56**%	**0.737**	**0.703**	**79.94**%	**0.714**	**0.695**	**79.18**%	**0.705**
AVA	v-SVR+	0.677	78.52%	0.665	0.662	76.43%	0.668	0.642	72.28%	0.651	0.582	68.84%	0.574
	RA-DCNN	0.547	74.62%	0.558	0.556	72.15%	0.534	0.538	70.11%	0.526	0.529	68.42%	0.517
	PI-DCNN	0.663	77.62%	0.676	0.638	73.48%	0.652	0.619	71.47%	0.628	0.592	69.93%	0.604
	ATTGAN	0.628	76.44%	0.642	0.608	74.18%	0.617	0.596	72.26%	0.604	0.587	70.04%	0.593
	SGAN	0.628	75.42%	0.635	0.614	73.71%	0.624	0.598	72.63%	0.605	0.584	69.18%	0.592
	SAGAN-	0.633	76.42%	0.638	0.614	74.68%	0.621	0.601	72.13%	0.598	0.587	70.07%	0.592
	SAGAN	**0.750**	**82.16%**	**0.758**	**0.734**	**81.03%**	**0.742**	**0.717**	**80.66%**	**0.722**	**0.695**	**79.42%**	**0.704**

TABLE III: Comparison to state-of-the-art attribute-aware methods on the AADB and AVA database. Among these methods, DCNN is supervised methods, while the others are semi-supervised methods.

Model	AADB			Model	AVA		
	SRCC	Accuracy	PLCC		SRCC	Accuracy	PLCC
RA-DCNN	0.678	78.62%	0.685	RA-DCNN	0.558	77.03%	0.563
DCNN	0.689	-	-	DCM	-	78.08%	-
MTDL	0.680	-	-	MTRLCNN	-	79.08%	-
ATTGAN	0.704	79.94%	0.708	ATTGAN	0.631	77.23%	0.655
PI-DCNN	0.705	81.05%	0.698	PI-DCNN	0.658	76.2%	0.672
SAGAN	**0.761**	**83.04**%	**0.766**	**SAGAN**	**0.774**	**83.72**%	**0.788**

B. Results of Within-Database Experiments

1) Analyses of Four Missing Rates: Table II demonstrates the within-database experimental results of semi-supervised attribute learning on the AADB and AVA databases. Among these methods, (i) v-SVR+ is the work using attributes on support vector regression; (ii) ATTGAN, SGAN and SAGAN- are the works of GAN variations; (iii) RA-DCNN and PI-DCNN are the works applying attributes on deep convolutional neural network.

From Table II, we make the following observations: First, compared to v-SVR+, RA-DCNN, PI-DCNN and ATTGAN, the proposed SAGAN has the best performance. For example, with the 40% missing rate, the performance of SAGAN are 7.5%/5.1%/7.6%, 9.2%/7.0%/10.8%, 9.3%/6.5%/8.9% and 9.0%/7.1%/13.9% higher SRCC/Accuracy/PLCC than v-SVR+, RA-DCNN, PI-DCNN and ATTGAN on the AADB database. Similarly significant improvement can be found on the AVA database. This demonstrates the superiority of the proposed method due to more attributes generated in SAGAN. Second, SAGAN performs better in all scenarios compared to SGAN that does not output aesthetic attributes on R network. The reason is that SAGAN learns the aesthetic attributes and score simultaneously to capture their dependencies and extract better feature representations. Third, when the missing rate varies from 10% to 40%, the gap performance between SAGAN and the other methods is becoming large. This is because the proposed method can generate more representative aesthetic attributes after missing, which is more beneficial for photo aesthetic assessment. Fourth, compared to SAGAN- that removes photo generator G from our method, SAGAN has a better performance in all scenarios. This is attributed to the fact that photo generator G leveraged to reconstruct the generated feature vector is important improvement for our proposed method.

2) Comparison of the Predicted Photo Examples: Figure 3 and Figure 4 show the predicted aesthetic scores of our methods compared to two attribute-aware methods (RA-DCNN and ATTGAN) and ground truth. We select some photos from the testing set of the AADB and AVA databases, respectively. Generally, we observe that our method has a lower prediction error compared to RA-DCNN and ATTGAN. The slight gap between our predicted scores and ground truth indicates that our trained models can accurately predict aesthetic scores.

3) Comparison to State-of-the-art Attribute-Aware Methods: To further evaluate the performance of the proposed method, we compare the proposed semi-supervised attributes learning method to related attribute-aware methods shown in Table III. From Table III, we find that on the AADB and AVA database, the proposed SAGAN method has the best performance w.r.t. the highest SRCC, accuracy and PLCC than the other methods. Compare to these methods using the same number of attributes that both databases provide, the proposed SAGAN method benefits both the score-attributes generator and photo generator tasks via generating informative signals. Compared to state-of-the-art attribute-aware methods that heavily rely on manual attributes annotation, the proposed method uses only partially attribute-annotated photos and achieves the best performance of aesthetic prediction.

4) Comparison to State-of-the-art Attribute-Unaware Methods: To further demonstrate the importance of high-level attributes in aesthetic assessment, we compare our method to state-of-the-art attribute-unaware methods shown in Table IV. As we can see from Table IV, the proposed method using high-level attribute in the aesthetic rating network achieves the best

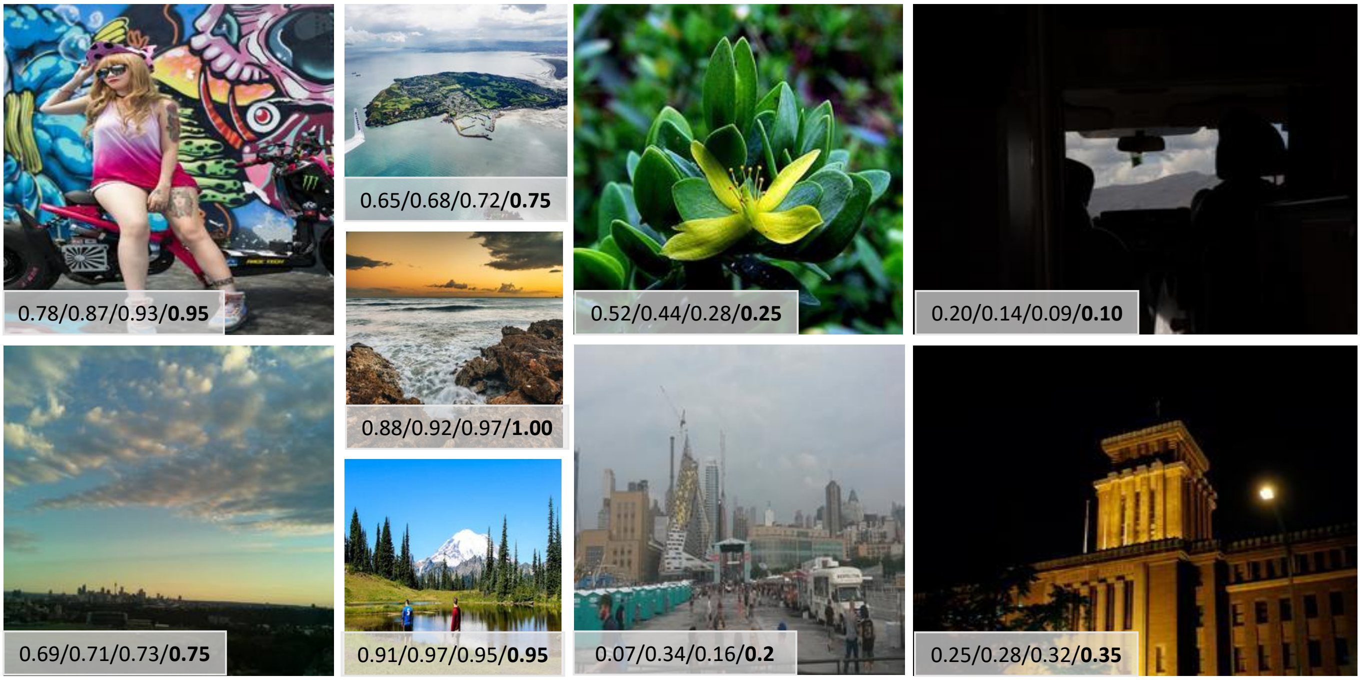

Fig. 3: Examples of predicted aesthetic scores on the AADB database: RA-DCNN/ATTGAN/Ours/Ground truth. The left two columns are images with high aesthetic scores and the remaining columns are images with low aesthetic scores.

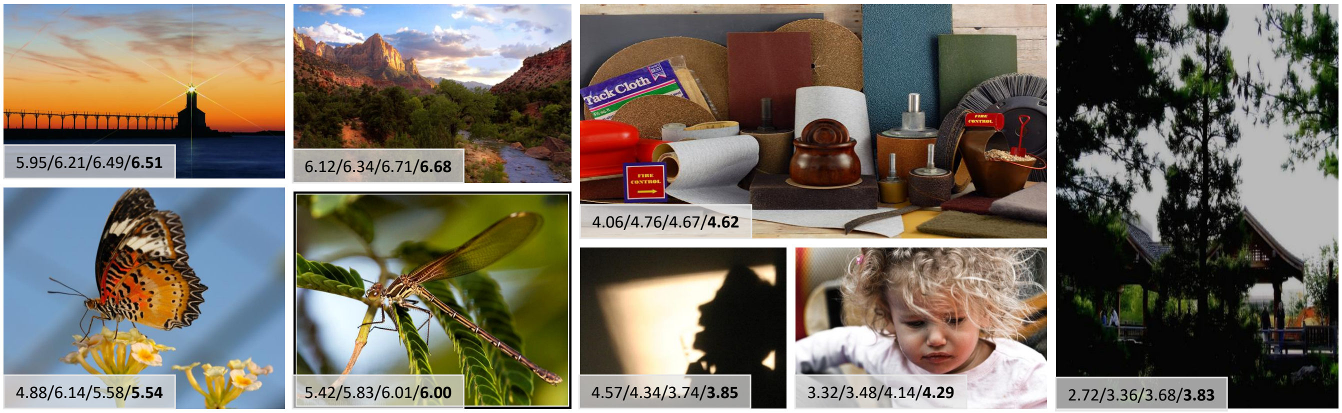

Fig. 4: Examples of predicted aesthetic scores on the AVA database: RA-DCNN/ATTGAN/Ours/Ground truth. The left two columns are images with high aesthetic scores and the remaining columns are images with low aesthetic scores.

TABLE IV: Comparison to state-of-the-art attribute-unaware methods on the AADB and AVA databases.

Model	AADB	Model	AVA		
	SRCC		SRCC	Accuracy	PLCC
Square-EMD	0.689	NIMA	0.612	-	0.636
AUPF	0.726	GPF-CNN	0.690	81.81%	0.704
CAN	0.710	MLSP	0.756	81.72%	0.757
CORF	0.677	CORF	0.671	-	0.665
Ours	**0.761**	**Ours**	**0.774**	**83.72%**	**0.788**

performance w.r.t. the highest SRCC, accuracy and PLCC than the other methods, which demonstrates that exploiting high-level aesthetic attribute can significantly benefit aesthetic assessment task and our proposed method successfully leverages the correlation between aesthetic attributes and aesthetic score.

C. Results of Cross-Database Experiments

1) Analyses of Four Missing Rates: The cross-database experimental results of semi-supervised adversarial learning for attribute-aware photo aesthetic assessment task on the AADB and AVA databases are shown in Table V. From Table V, we find: First, when the methods are trained on the AADB database and tested on the AVA database, the proposed SAGAN achieves the best performance in the different missing rates of attributes. For example, with the 40% missing rate, the performance of SAGAN are 8.8%/7.8%/7.6%, 13.7%/9.15%/12.8%, 7.5%/4.8%/6.1% and 6.8%/3.4%/5.4% higher SRCC/Accuracy/PLCC than the method of v-SVR+, RA-DCNN, PI-DCNN and ATTGAN. Second, the SGAN perform better than RA-DCNN in all cases. Although SGAN

TABLE V: Cross-database evaluation of semi-supervised learning for aesthetic assessment with four missing rates of attribute-annotated photos.

Database	Methods	10%			20%			30%			40%		
		SRCC	Accuracy	PLCC	SRCC	Accuracy	PLCC	SRCC	Accuracy	PLCC	SRCC	Accuracy	PLCC
From AADB to AVA	v-SVR+	0.254	59.88%	0.261	0.232	58.43%	0.248	0.206	55.42%	0.217	0.192	53.39%	0.202
	RA-DCNN	0.182	57.09%	0.197	0.173	55.62%	0.179	0.154	53.37%	0.162	0.143	52.03%	0.150
	PI-DCNN	0.274	60.57%	0.271	0.253	59.98%	0.248	0.221	57.72%	0.236	0.205	56.43%	0.217
	ATTGAN	0.263	60.02%	0.270	0.241	59.34%	0.254	0.226	58.52%	0.238	0.217	57.79%	0.224
	SGAN	0.243	58.77%	0.250	0.232	56.40%	0.244	0.225	55.38%	0.228	0.210	55.02%	0.214
	SAGAN-	0.259	59.63%	0.263	0.234	58.02%	0.240	0.225	57.36%	0.232	0.209	56.48%	0.205
	SAGAN	**0.327**	**68.82%**	**0.334**	**0.306**	**66.47%**	**0.312**	**0.288**	**62.32%**	**0.285**	**0.280**	**61.18%**	**0.278**
From AVA to AADB	v-SVR+	0.486	68.94%	0.498	0.463	66.43%	0.471	0.445	64.12%	0.450	0.417	62.39%	0.412
	RA-DCNN	0.332	67.54%	0.340	0.313	64.39%	0.311	0.298	60.12%	0.304	0.283	59.33%	0.293
	PI-DCNN	0.493	69.42%	0.502	0.477	67.99%	0.480	0.442	64.33%	0.447	0.394	63.50%	0.405
	ATTGAN	0.506	70.62%	0.511	0.488	68.54%	0.492	0.463	66.12%	0.268	0.432	63.12%	0.444
	SGAN	0.492	69.11%	0.498	0.477	68.12%	0.482	0.453	66.53%	0.451	0.442	64.88%	0.437
	SAGAN-	0.502	70.58%	0.506	0.487	69.02%	0.492	0.465	67.66%	0.472	0.450	65.32%	0.448
	SAGAN	**0.583**	**74.43%**	**0.601**	**0.565**	**72.12%**	**0.577**	**0.554**	**70.05%**	**0.561**	**0.547**	**69.32%**	**0.552**

is not assisted by attributes in R network, it takes advantage of the competition between photo synthesis and aesthetic score assessment, hence still achieves the better performance than RA-DCNN. Third, compared to the method of SAGAN- that removes photo generator G, the method of SAGAN perform better, which demonstrate photo generator plays an important roles in regularizing the score-attributes generator. A similar conclusion can be found on the AVA database.

TABLE VI: Cross-database evaluation of SRCC reported on current studies. Among these four methods, RA-DCNN and SAGAN are semi-supervised methods while NIMA and AUPF are attribute-unaware methods.

Model	From AADB to AVA	From AVA to AADB
NIMA	0.2950	0.4732
AUPF	0.3225	0.5176
RA-DCNN	0.1566	0.3191
SAGAN	**0.3462**	**0.6052**

2) Comparison to the State-of-the-art Methods: Table VI shows the related cross-database experimental results on the AADB and AVA databases. Compared to current studies on cross database, the proposed method has a better performance when training on the AADB testing on the AVA and training on the AVA testing on the AADB. This demonstrates that the proposed method further explores the adaptation of different photo groups and photo collections. Note that cross-database performance is worse than within-database performance because the AADB database contains photos with different distributions of visual characteristics compared to the AVA database. Photos on the AVA database are professionally photographed or heavily edited; while many photos in the AADB database are related to many daily life. Thus, the generalization of the proposed method should be further improved in the future work.

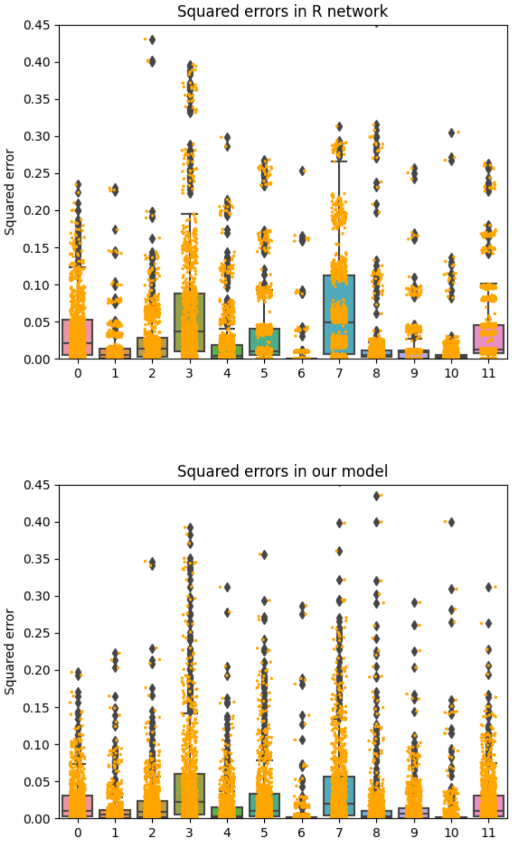

Fig. 5: Visualization of squared error between predicted results and ground truth with 40% missing rate of attributes on the AADB database. Digit $0-11$ in x-axis represents score and 11 attributes, respectively.

D. Ablation Study

1) Analysis of Adversarial Learning: This section aims to evaluate how the proposed adversarial learning is effective for aesthetic score and attributes generation. Note that our method without the adversarial learning turns into the remaining R network and G network. We compare the remaining networks with our original method, and visualize their squared errors of predictions and ground truth on the AADB database with 40% missing rate. In Figure 5, we depict 0-11 columns in the x-axis corresponding to the results of one score and 11 attributes. The orange points in each column represent their squared errors. The boxes behind the orange points depict the ranges of their corresponding values. Compared with the top figure, the box height in the bottom figure becomes lower and orange points are more merged when the generative network and adversarial network are involved in our method. This indicates that the adversarial learning in our method can be used to explore the dependencies among photo features, aesthetic attributes and aesthetic score, which thus can improve the performance of photo aesthetic prediction.

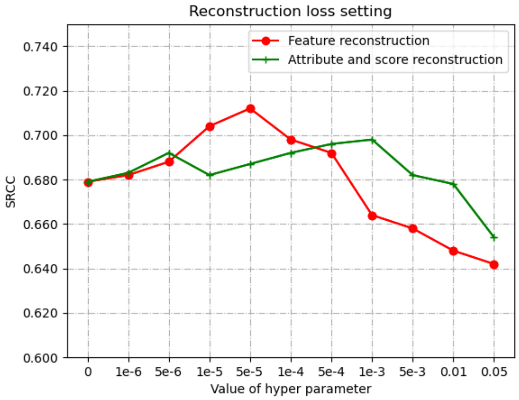

Fig. 6: Our performance with respect to feature reconstruction (λ_{rg}) and score-attributes reconstruction (λ_{rr}) on the AVA database.

2) Evaluation of Reconstruction Loss: As elaborated in Section IV-C, reconstruction loss is designed in the process of feature generation and score-attributes generation. It is vital to help the G network to restore photo features and help R network to generate photo attributes and score. To explore the impact of reconstruction loss, we conduct experiments by setting different values of λ_{rg} and λ_{rr}. Take the AVA database for example, our experimental performances w.r.t λ_{rg} and λ_{rr} are shown in Figure 6. When λ_{rg} varies, we set $\lambda_{rr} = 0$ and vice versa. From Figure 6, the SRCC curves significantly increase then decline followed by the change of λ_{rg} and λ_{rr}. Thus, we can find the optimal value of λ_{rg} and λ_{rr}.

VI. Conclusion

In this paper, we propose a semi-supervised learning method from partially attribute-annotated photos with the aim of reducing the reliance on manual attributes annotation for aesthetic assessment. An adversarial training framework is proposed to explore the joint distribution among photo features, aesthetic attributes and aesthetic score. Aesthetic attributes are used as privileged information to construct a better score-attribute generator and a photo generator. Supervised losses are used for the proposed networks predicting on commonly occurring range. Reconstruction losses are designed for regularizing the score-attributes generation and photo generation. Experimental results on the AADB and AVA databases demonstrate that the proposed method successfully leverages the inherent connections between aesthetic attributes and aesthetics through semi-supervised attributes learning and adversarial training process, and thus achieves the superior performance.

References

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

[2] P. Lv, M. Wang, Y. Xu, Z. Peng, J. Sun, S. Su, B. Zhou, and M. Xu, "Usar: An interactive user-specific aesthetic ranking framework for images," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1328–1336.

[3] V. Hosu, B. Goldlucke, and D. Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9375–9383.

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

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

[6] T. O. Aydın, A. Smolic, and M. Gross, "Automated aesthetic analysis of photographic images," *IEEE transactions on visualization and computer graphics*, vol. 21, no. 1, pp. 31–42, 2015.

[7] M. Fan, R. Huang, W. Feng, and J. Sun, "Image blur classification and blur usefulness assessment," in *Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on*. IEEE, 2017, pp. 531–536.

[8] S. Herron, "Technology of duotone color transformations in a color managed workflow," in *Color Imaging VIII: Processing, Hardcopy, and Applications*, vol. 5008. International Society for Optics and Photonics, 2003, pp. 365–371.

[9] H. Wakabayashi, K. Kazami, T. Sosa, and H. Miyamoto, "Camera having soft focus filter," Jun. 26 1990, uS Patent 4,937,609.

[10] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma, "Dual learning for machine translation," *Advances in neural information processing systems*, vol. 29, pp. 820–828, 2016.

[11] H. Yang, P. Shi, S. He, D. Pan, Z. Ying, and L. Lei, "A comprehensive survey on image aesthetic quality assessment," in *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*. IEEE Computer Society, 2019, pp. 294–299.

[12] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017.

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

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

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

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

[17] Y. Kao, R. He, and K. Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1482–1495, 2017.

[18] C. Cui, H. Liu, T. Lian, L. Nie, L. Zhu, and Y. Yin, "Distribution-oriented aesthetics assessment with semantic-aware hybrid network," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1209–1220, 2018.

[19] G. Malu, R. S. Bapi, and B. Indurkhya, "Learning photography aesthetics with deep cnns," *arXiv preprint arXiv:1707.03981*, 2017.

[20] Y. Shu, Q. Li, C. Xu, S. Liu, and G. Xu, "V-svr+: Support vector regression with variational privileged information," *IEEE Transactions on Multimedia*, 2021.

[21] B. Pan, S. Wang, and Q. Jiang, "Image aesthetic assessment assisted by attributes through adversarial learning," in *AAAI 2018*, vol. 33, 2019, pp. 679–686.

[22] W.-S. Lai, J.-B. Huang, and M.-H. Yang, "Semi-supervised learning for optical flow with generative adversarial networks," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 353–363.

[23] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[24] Z. Wang, Q. Xu, K. Ma, Y. Jiang, X. Cao, and Q. Huang, "Adversarial preference learning with pairwise comparisons," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 656–664.

[25] J. Dong and T. Lin, "Margingan: Adversarial training in semi-supervised learning," *Advances in neural information processing systems*, vol. 32, pp. 10 440–10 449, 2019.

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

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

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] D. P. K. JLB, "Adam: A method for stochastic optimization," in *3rd international conference for learning representations, San Diego*, 2015.

[30] L. Li, H. Zhu, S. Zhao, G. Ding, H. Jiang, and A. Tan, "Personality driven multi-task learning for image aesthetic assessment," in *ICME 2019*. IEEE, 2019, pp. 430–435.

[31] Y. Shu, Q. Li, S. Liu, and G. Xu, "Learning with privileged information for photo aesthetic assessment," *Neurocomputing*, 2020.

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

[33] L. Hou, C.-P. Yu, and D. Samaras, "Squared earth mover's distance-based loss for training deep neural networks," *arXiv preprint arXiv:1611.05916*, 2016.

[34] H. Zeng, Z. Cao, L. Zhang, and A. C. Bovik, "A unified probabilistic formulation of image aesthetic assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 1548–1561, 2019.

[35] D. Liu, R. Puri, N. Kamath, and S. Bhattacharya, "Composition-aware image aesthetics assessment," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3569–3578.

[36] H. Zhu, H. Shan, Y. Zhang, L. Che, X. Xu, J. Zhang, J. Shi, and F.-Y. Wang, "Convolutional ordinal regression forest for image ordinal estimation," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[37] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[38] X. Zhang, X. Gao, W. Lu, and L. He, "A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction," *IEEE Transactions on Multimedia*, 2019.

[39] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural networks*, vol. 22, no. 5-6, pp. 544–557, 2009.

Yangyang Shu received his B.S. degree in computer science from Anhui University in 2015, and received his M.S. degree in Computer Science in the University of Science and Technology of China in 2018. Now he is currently pursuing his ph.D degree in Faculty of Engineering and Information Technology in the University of Technology Sydney, Australia. His research interests lie in machine learning, pattern recognition, multimedia, privileged information and related applications in artificial intelligence, including multi-label learning, music emotion recognition and photo aesthetic assessment.

Qian Li has been a Lecturer at the School of Engineering, Computing and Mathematical Sciences (EECMS), Curtin University, Australia. She received her Ph.D. in Computer Science from the Chinese Academy of Science. Her general research interests lie primarily in optimization algorithms, topological data analysis, and causal machine learning. Her papers have been published in the top-tier conferences and journals in the field of machine learning and computer vision.

Lingqiao Liu is a senior lecturer at the University of Adelaide and the Australian Institute for Machine Learning. He received his BS and MS degrees in communication engineering from the University of Electronic Science and Technology of China, Chengdu, in 2006 and 2009, respectively, and the Ph.D. degree from the Australian National University, Canberra, in 2014. In 2016, he was awarded the Discovery Early Career Researcher Award from the Australian Research Council and the University Research Fellow from the university of Adelaide. His current research interests include low-supervision learning and various topics in computer vision and natural language processing.

Guandong Xu is currently a Full Professor with the School of Computer Science, University of Technology Sydney, Ultimo, NSW, Australia. He has published 3 monographs in Springer and CRC press and more than 250 journal articles and conference papers in data science, recommender systems, text mining, and social network analysis. He has published three monographs in Springer and CRC press, and 180+ journal and conference papers including TOIS, TIST, TNNLS, TSC, TIFS, IEEE-IS, Inf. Sci., KAIS, WWWJ, KBS, Neurocomputing, ESWA, Inf. Retr., IJCAI, AAAI, WWW, KDD, ICDM, ICDE, CIKM. Dr. Xu has served as a Guest Editor for Pattern Recognition, the IEEE Transactions on Computational Social Systems, Journal of Software and Systems, and World Wide Web journal etc. He is Assistant Editor-in-Chief of World Wide Web journal.