Aesthetic Attribute Assessment of Images Numerically on Mixed Multi-attribute Datasets

XIN JIN, XINNING LI, HAO LOU, CHENYU FAN, QIANG DENG, and CHAOEN XIAO,
Beijing Electronic Science and Technology Institute, China
SHUAI CUI, University of California, Davis, USA
AMIT KUMAR SINGH, National Institute of Technology Patna, India

With the continuous development of social software and multimedia technology, images have become a kind of important carrier for spreading information and socializing. How to evaluate an image comprehensively has become the focus of recent researches. The traditional image aesthetic assessment methods often adopt single numerical overall assessment scores, which has certain subjectivity and can no longer meet the higher aesthetic requirements. In this article, we construct an new image attribute dataset called aesthetic mixed dataset with attributes (AMD-A) and design external attribute features for fusion. Besides, we propose an efficient method for image aesthetic attribute assessment on mixed multi-attribute dataset and construct a multitasking network architecture by using the EfficientNet-B0 as the backbone network. Our model can achieve aesthetic classification, overall scoring, and attribute scoring. In each sub-network, we improve the feature extraction through ECA channel attention module. As for the final overall scoring, we adopt the idea of the teacher-student network and use the classification sub-network to guide the aesthetic overall fine-grain regression. Experimental results, using the MindSpore, show that our proposed method can effectively improve the performance of the aesthetic overall and attribute assessment.

CCS Concepts: • Applied computing → Media arts;

Additional Key Words and Phrases: Aesthetic mixed dataset with attributes, multitasking, external attribute features, ECA channel attention

ACM Reference format:
Xin Jin, Xinning Li, Hao Lou, Chenyu Fan, Qiang Deng, Chaoen Xiao, Shuai Cui, and Amit Kumar Singh. 2023. Aesthetic Attribute Assessment of Images Numerically on Mixed Multi-attribute Datasets. ACM Trans. Multimedia Comput. Commun. Appl. 18, 3s, Article 138 (March 2023), 16 pages.
https://doi.org/10.1145/3547144

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
1551-6857/2023/03-ART138 $15.00
https://doi.org/10.1145/3547144
1 INTRODUCTION

The visual aesthetic quality of the image measures the visual attraction of the images for humans. Since visual aesthetics is a subjective attribute [13], it always depends on personal emotions and preferences. This makes image aesthetic quality assessment a subjective task and evaluated only by experts. If there is a large number of image samples, then the efficiency of artificial aesthetic quality assessment will be quite low. However, people tend to agree that some images do indeed more attractive than others in daily life, which creates a computational aesthetics [5]. Computational aesthetics lets the computer mimic the process of human aesthetic assessment and compute the methods to predict the aesthetic quality of the images automatically.

The focus of computational aesthetics is to predict people’s emotional responses to visual stimuli through computational technology, study the internal mechanism of human perception, and explore the mystery of artificial aesthetic intelligence. Computational visual aesthetics [3] is the computational processing of human visual information. Image aesthetic assessment is the most popular research direction in the field of computational visual aesthetics, and it is also the first step in studying computational visual aesthetics. Image aesthetic assessment is to simulate human perception and cognition by the computers, which can provide the quantitative assessment of aesthetic quality. Image aesthetic attributes assessment mainly focuses on the quantitative assessment formed by aesthetic attributes such as composition, color, and light in the photographic images.

There are two main parts for the image aesthetics quality assessment: the feature extraction part and the assessment part. In the feature extraction, Yan et al. [16] proposed the 7-dimensional aesthetic features based on photography knowledge and high-level semantics. The features include simplicity, spatial edges distribution, color distribution, hue count, blur, contrast, and brightness. With the development of deep learning, researchers introduce deep convolutional neural networks in the task of image aesthetics assessment. Due to the ability of learning features automatically, people can extract aesthetic features from images by the deep convolutional neural networks without a lot of aesthetic knowledge and professionally photography experiences. However, there are deficiencies in general aesthetic benchmark datasets. For example, in AADB [17], each image is only marked by a small number of annotation. So, the data labels is kind of subjective. It is difficult to extract the aesthetic attribute features and limits the extraction ability of features for the model. At the same time, the multidimensional attribute assessment will lead to a sharp increase in the number of network parameters, which is not conducive to the actual development and application.

To solve these problems, we propose an efficient method for image aesthetic attribute assessment based on the mixed multi-attribute dataset. Our work includes the following three contributions:

(1) We screen and construct the aesthetic mixed dataset with attributes (AMD-A), which has more image aesthetic attribute annotations compared with the traditional aesthetic attribute datasets.

(2) We propose an aesthetic multitasking network architecture based on EfficientNet-B0 and ECA channel attention modules to simplify the model parameters and realize the aesthetic classification, overall scoring, and attribute scoring.

(3) We design several kinds of external attribute features and use feature fusion to improve the performance of the image aesthetic attribute assessment. Besides, inspired by teacher-student network, we propose the soft loss to improve the performance of the image aesthetic overall assessment.
Fig. 1. The examples for image aesthetic attribute assessment. S represents the overall score, C represents the color score, L represents the light score, and CM represents the composition score.

2 RELATED WORK

2.1 Image Aesthetic Assessment

Image aesthetic assessment is usually regarded as a classification or score regression task. For the classification task, images can generally be divided into high-quality images and low-quality images [31]. For the regression task, images can be evaluated according to aesthetic overall scores. However, the overall scores cannot accurately measure the results of aesthetic assessment, which largely ignores the diversity, subjectivity, and personality in the human aesthetic consensus. An image also contains many aesthetic attributes such as light and shadow, color, composition, blur, movement, and interest. So, the related work has focused on the multidimensional aesthetic attribute assessment (see Figure 1).

In the early stage, the handcraft designing features was the main way to extract features from images [23, 25, 28, 33]. Datta et al. [5] used bottom-level features (color, texture, shape, image size, etc.) and high-level features (depth of field, rule of thirds, regional contrast) as the image aesthetic features. Marchesotti [25] and others directly used SIFT (BOV or FisherVector) and local color descriptors to classify aesthetic images.

Today, deep learning has been widely used in different walks of life, such as IoV [34], IoT [35], and edge intelligence [37]. With the development of deep learning, the researchers introduce deep convolutional neural networks into the image aesthetic assessment. Due to the powerful learning capabilities in deep convolutional neural networks, people can automatically extract aesthetic features without substantial theoretical knowledge and photographic experience. In recent years, deep convolutional neural networks have shown good performance in the overall regression and classification tasks of image aesthetics, and a series of excellent models [6, 15, 24, 30, 36, 38, 41] have emerged. At the same time, in the field of image aesthetic attribute assessment, there are image aesthetic attribute scoring methods based on hierarchical multitasking networks [11] and incremental learning multitasking networks [12] by using image composition, light and color, exposure depth of field, and other attributes. However, due to the limitation of data quantity and the categories of aesthetic attributes, the above attribute assessment methods do not have high accuracy. With the increasing of attribute categories, the network models become extremely complex, which is not conducive to expansion and practical application.

2.2 Aesthetic Attribute Datasets

The emergence of large-scale aesthetic datasets provides a rich source of samples for aesthetic assessment model training. Murray et al. [27] constructed an aesthetic visual analysis dataset (AVA) containing 255,530 images from the www.dpchallenge.com. AVA is the benchmark dataset for the current image aesthetic quality assessment task. Each image includes a numerical aesthetic overall regression label, 66 semantic labels, and 14 style labels. Building on this dataset, Kang et al. [14] proposed the explainable visual aesthetics dataset (EVA), which contains 4,070 images. Each image has 30 annotation scores at least and mainly includes an overall score and four...
different aesthetic attribute scores. Compared with AVA, EVA adopts a more rigorous annotation collection method and overcomes the noisy labels due to the misunderstanding bias of the annotators, which facilitates the research on aesthetic understanding.

Kong et al. [17] designed the **aesthetics and attributes database (AADB)**. The dataset contains 9,958 images from professional photographers and ordinary photographers, and each image has an overall score and 11 aesthetic attributes scores. Chang et al. [4] first proposed the annotation information for aesthetic language assessment and designed the **photo critique captioning dataset (PCCD)**, which contains 4,235 valid images from the foreign photography website Gu-rushots.com. Except for the overall score, each image has language comments and scores on the composition perspective, color illumination, image theme, depth of field, focus, and camera use. Fang et al. [7] conducted the first systematic study on the smartphone image aesthetic quality assessment and constructed **smartphone photography attribute and quality dataset (SPAQ)**. The dataset consists of 11,125 images taken by 66 smartphones, and each image including an overall label, an image attribute label, and a scene category label.

2.3 Feature Fusion

Feature fusion is a common way to improve the model performance. In the field of aesthetic assessment, the low-level features obtained by deep neural networks have high resolution and contain more aesthetic low-level information, such as color, texture, and structure. But after fewer convolutional layers, low-level features contain lower high-level semantics and are easy to cause noise interference [39, 40]. High-level features tend to have stronger semantic information, but they have very low feature resolution and poor perception of details. In the aesthetic field, local aesthetic attributes based on low-level information are as important as global aesthetic features based on high-level information, especially in the regression task of aesthetic attributes. The output results of deep convolutional networks are often in the last layer of the network, which leads to the inability to effectively extract the corresponding attribute features in the assessment of aesthetic attributes.

In related work, the researchers improve the performance of detecting and dividing objects by integrating the network layer features at different locations of the neural network. We classified feature fusion and predicted output as early and late fusion. The idea of early fusion strategy is to integrate the high-level and low-level features and then conduct model training and prediction on the fusion features. Such methods usually use concatenating and adding operations to fuse the features. Related research work such as Inside-Outside Net [2] and HyperNet [18]. The idea of the late fusion strategy is to output the prediction results of different network layers and then to fuse all the detection results. Related research work such as Feature Pyramid Network [20], Single Shot MultiBox Detector [22], and Densenet [10]. This article mainly adopts the early fusion strategy to integrate different high-level features of the sub-networks and external attribute features to improve the supervised assessment performance of image aesthetic attributes.

3 AESTHETIC MIXED DATASET WITH ATTRIBUTES (AMD-A)

To construct a dataset with a reasonable distribution both in image aesthetics quality assessment and image aesthetic attributes assessment, we rebuild a dataset named **Aesthetic mixed dataset with attributes (AMD-A)** concluding 16,924 images. According to different tasks, AMD-A is divided into two sets. One set (11,166 images) is applied to aesthetic overall score regression, another set (16,924 images) is applied to aesthetic attribute score classification and regression.

As for the attribute regression, we collect images from EVA [14], AADB [17], PCCD [4], PADB, and HADB. PADB and HADB are self-built datasets. Each image has three attribute labels and one overall label for assessment. The three attributes include light, color, and composition. To increase
Aesthetic Attribute Assessment of Images Numerically

Table 1. Means and Standard Deviations for Light, Color, Composition, and Overall Labels

|                  | light  | color | composition | overall |
|------------------|--------|-------|-------------|---------|
| Mean             | 0.5324 | 0.5472| 0.5420      | 0.5398  |
| Standard deviation | 0.1522 | 0.1319| 0.1388      | 0.2116  |

Fig. 2. The distribution of light, color, composition, and overall labels in AMD-A.

Fig. 3. The architecture of the network.

the training samples of the aesthetic regression, we collected 5,758 images with only overall scores from other datasets. There are from DPCallenge.com, SCUI-FBP5500 [19], Photo.net and SPAQ [7]. Data labels are continuous and each one has a score range of 0–1. The distribution of different labels are as shown in Figure 2.

We calculated the means and the standard deviations of the labels in Table 1. The results show that the four different categories of labels have the similar data distributions.

4 NETWORK ARCHITECTURE

We build a multitasking network architecture including the backbone network and five sub-networks. The basic network architecture is shown in Figure 3.

We use EfficientNet-B0 as the backbone network [32]. EfficientNet has an efficient feature extraction ability and can achieve more accurate aesthetic regression with a small number of parameters. The size of the parameter model for EfficientNet-B0 is only 20M, which is smaller
than many existing model parameter models. So, it is an ideal pre-trained parameter model in development and application scenarios. We use the blue part for the backbone network in Figure 3.

Between the backbone network and the sub-networks, there are ECA channel attention modules. In ECA, GAP is the global average pooling, $\sigma$ is the sigmoid function, $\psi(C)$ is the function of parameter $k$, and each layer will be activated by the relu function. We will introduce it in Section 4.1.

The five sub-networks are one classification sub-network, three attribute sub-networks, and one regression sub-network consisting of fully connected layers. We use the red part for the classification sub-network, the khaki part for the attribute sub-networks, and the green part for the regression sub-network. The attribute sub-networks in the figure have the similar structure, so we omitted additional attribute sub-networks and displayed only one attribute sub-network. The yellow part is for the external attribute features (4 light features, 7 color features, and 10 composition features).

From the architecture of the network, we can see two different kinds of feature fusions. One for external features and 10 attribute features in the attribute regression tasks and another one for 10 regression features and 30 attribute features in the regression task. The training details will be explained in Section 4.2.

### 4.1 ECA Channel Attention

From the network structure in Figure 3, each ECA channel attention module is added between the backbone network and each sub-network. Recently, there are a lot of researches improving channel and spatial attention to make progresses in the performance. The ECA channel attention can improve the ability of extracting features. It mainly improves from the SENet [29] module. ECA channel attention proposes a local cross-channel interaction strategy without the dimension reduction and implements a method of adaptively selecting the size of the one-dimensional convolution kernel. So, we use it to improve the performance of the image aesthetic attribute and overall assessment. To the best of our knowledge, it is also the first time the ECA channel attention is used in the task of aesthetics attribute assessment.

Instead of reducing global average pooling layers at the channel level, ECA channel attention captures local cross-channel interaction information by considering each channel and its $k$ neighbors. This can be effectively achieved efficiently by a fast one-dimensional convolution. $k$ indicates how many neighbors near the channel participate in the attention prediction of this channel. To avoid artificially raising $k$ by cross-validation, Reference [29] proposes a way to determine $k$ automatically. The size of the convolutional kernel $k$ is proportional to the channel dimension as follows:

$$k = \psi(C) = \frac{\frac{\log_2(C)}{Y} + b}{Y_{\text{odd}}}.$$  \hspace{1cm} (1)

$|t|_{\text{odd}}$ means the nearest odd number of $t$, the value of $\gamma$ is 2, the value of $b$ is 1. We can get $k = 7$ if $C = 1,280$.

### 4.2 Multitasking Network

We use EfficientNet-B0 as the backbone network and design the classification sub-network for the classification task. Between these two networks, we use the ECA channel attention module to improve the capability of the aesthetic feature extraction. The classification task divides aesthetic images into 10 categories based on the level of aesthetic scores, which is a coarse-grained aesthetic regression. During training, we relaxed the parameters of the backbone network and the classification sub-network, obtained the high-level aesthetic features of $1,280 \times 7 \times 7$ through the EfficientNet-B0, and made the feature reweighted through the ECA channel attention module. After that, we used the global average pooling layer to transform the high-level features into
After completing the classification task, we trained the three sub-networks in the order of light, color, and composition. The training labels are the attribute labels of each image. The structure of attribute sub-networks are similar to the classification sub-networks. We reduced 1,280 features to 10 features by one fully connected layer and the activation function. These 10 features are the attribute features of the image. As shown in Figure 4, we spliced different external features onto attribute features and regress the features to an attribute score using one fully connected layer and the activation function. The design and calculation of the external features will be mentioned in Section 5.

$$\text{Loss} = \text{MSELoss} + \lambda \text{SoftLoss}$$  \hspace{1cm} (2)

As shown in Figure 5, we also reduced 1,280 features to 10 features in the regression sub-network. Then, we combined these 10 features with $3 \times 10$ attribute features in the penultimate layer of the attribute sub-network to obtain an aesthetic overall score through one fully connected layer and the activation function. Inspired by the idea of teacher-student networks [9], we used 10 features from the output of the classification sub-network to guide the 10 features of the regression sub-network. We calculate the relative entropy between the two feature sequences and define it as the soft loss. The regression sub-network is gradient-adjusted by the MSE loss of the aesthetic score regression and a weighted soft loss. We set the weight of the soft loss as 0.1, which we will prove in Section 6.

5 ATTRIBUTE FEATURES FUSION

Before training the three attribute regression sub-networks, we will design and calculate the external attribute features of the input images. The external attribute features includes light features, color features, and composition features. We will store the external attribute features in the data labels. The design details are presented in this section.

5.1 Light Features

Referring to the value and lightness information [5], we extract the light features as mean of value ($f_1$), standard deviation of value ($f_2$), mean of lightness ($f_3$), and standard deviation of lightness ($f_4$).

The light features was mainly obtained by calculating the mean and standard deviation of the L and V channels. Both the value and lightness range is 0–255. We use $x$ to represent the pixels in
Table 2. Light Feature Values for Samples

| Light feature                  | 6(a)        | 6(b)        | 6(c)        | 6(d)        |
|--------------------------------|-------------|-------------|-------------|-------------|
| Mean of value ($f_1$)          | 148.253016  | 34.985529   | 72.411960   | 139.248896  |
| Standard deviation of value ($f_2$) | 52.104682  | 40.385166   | 29.037064   | 39.939841   |
| Mean of lightness ($f_3$)       | 183.823056  | 148.253016  | 79.659121   | 151.227269  |
| Standard deviation of lightness ($f_4$) | 50.643524  | 78.682690   | 28.775265   | 43.909947   |

The calculation formulas of average brightness ($f_1$), standard deviation of brightness ($f_2$), average lightness ($f_3$), and standard deviation of lightness ($f_4$) are as follows:

$$f_1 = \frac{1}{|I|} \sum_{x \in I} L(x),$$  \hspace{1cm} (3)

$$f_2 = \text{std}(L(x)),$$  \hspace{1cm} (4)

$$f_3 = \frac{1}{|I|} \sum_{x \in I} V(x),$$  \hspace{1cm} (5)

$$f_4 = \text{std}(V(x)).$$  \hspace{1cm} (6)

Figure 6 shows four sample images of different light effects, where Figures 6(a) and 6(b) have better light effects, while Figures 6(c) and 6(d) have poor light effect. Table 2 shows the light feature values corresponding to these four images. From the feature values in Table 2, we can see that the images with better light effect generally have higher means and standard deviations. While the image is too dark, the means and standard deviations are lower, and if the exposure is too excessive, the means will be higher and the standard deviations will be lower.

5.2 Color Features

Referring to the color information [1], we extract the color features as the weight of color channel ($f_1$), the number of RGB dominant colors ($f_2$), the degree of RGB dominant colors ($f_3$), the number of HSV dominant colors ($f_4$), the degree of HSV dominant colors ($f_5$), the number of dominant hues ($f_6$), the contrast ratio of dominant hues ($f_7$).
By calculating the approximation degree of the RGB color channel, the images can be divided into colored three-channel images, approximate grayscale channel images, and grayscale single-channel images. The weight of color channel \( f_1 \) is calculated by converting the image into the RGB channels. If the values of the three channel are exactly the same, then \( f_1 = 0 \). If the three channels are not exactly consistent, then we will calculate the average difference between the RGB channels at each pixel, while the average difference is less than 10, \( f_1 = 0.5 \); otherwise, the images are considered as the colored three-channel images, and \( f_1 = 1 \).

The dominant colors are mainly calculated by the color histogram in the RGB channel and the hue histogram in the HSV channel. As for RGB dominant colors, we quantize each RGB channel to 8 values, so a 512-dimensional histogram \( h_{RGB} = \{ h_0, h_1, \ldots, h_{511} \} \) can be created. \( h_i \) represents the number of pixels in the \( i \)th histogram. \( c_1 \) is the threshold parameter of the color number in the formula and \( c_1 = 0.01 \), which means if the number of the color in RGB is more then 51, then we can consider it as the RGB dominant color. We define the number of RGB dominant colors \( f_2 \) as follows:

\[
f_2 = \sum_{k=0}^{512} 1 \left( h_k \geq c_1 \max h_i \right).
\] (7)

The degree of RGB dominant colors \( f_3 \) indicates how the dominant color is occupied in the image colors. The formula is as follows:

\[
f_3 = \frac{\max_i h_i}{|I|}.
\] (8)

Similarly, replacing the RGB channel with the HSV channel, we can get the number of HSV dominant colors \( f_4 \) and the degree of HSV dominant colors \( f_5 \).

For \( f_6 \) and \( f_7 \), we eliminated pixels with saturation elimination values less than 0.2; in other words, we eliminated all light or dark pixels. We then calculate the hue histograms of the remaining pixels with 20 uniform intervals, each occupying 18° sectors of the hue ring. In the hue histogram \( H_{hue} = \{ h_0, h_1, \ldots, h_{19} \} \), \( h_i \) represents the pixel histogram in the \( i \)th interval. The formula for the number of dominant hues is as follows:

\[
f_6 = \sum_{i=1}^{20} 1 \left( |h_i| \geq c_2 |I| \right).
\] (9)

Similarly, we set \( c_2 = 0.01 \). The contrast ratio of dominant hues \( f_7 \) indicates the maximum contrast between the two principal hues in the image; the formula is as follows:

\[
f_7 = \max_{i,j} ||h_i| - |h_j||
\] (10)

Figure 7 shows four sample images of different color effects, where Figures 7(a) and 7(b) have better color effects, while Figures 7(c) and 7(d) have poor color effect. Table 3 shows the color feature values corresponding to these four images. From the feature values in Table 3, we can
Table 3. Calculation Results of Color Feature Sample

| Color feature                             | 7(a) | 7(b) | 7(c) | 7(d) |
|-------------------------------------------|------|------|------|------|
| Weight of color channel $f_1$             | 1.0  | 1.0  | 1.0  | 1.0  |
| Number of RGB dominant colors $f_2$       | 102  | 100  | 19   | 14   |
| Degree of RGB dominant colors $f_3$       | 0.087823 | 0.075602 | 0.253429 | 0.228055 |
| Number of HSV dominant colors $f_4$       | 220  | 171  | 49   | 39   |
| Degree of HSV dominant colors $f_5$       | 0.040854 | 0.052193 | 0.160361 | 0.197352 |
| Number of dominant hues $f_6$             | 8    | 10   | 5    | 4    |
| Contrast ratio of dominant hues $f_7$     | 162  | 162  | 18   | 18   |

see that the images with better color effect generally have more RGB and HSV dominant colors. Besides, the degree and the contrast ratio is higher. So, we can find that the images have the better color effects when they have richer colors.

5.3 Composition Features

We extract the composition features as golden section ($f_1$), center ($f_2$), slant ($f_3$), the triangle ($f_4$), guideline ($f_5$), rule of thirds ($f_6$), symmetry ($f_7$), diagonal ($f_8$), frame ($f_9$), and circle ($f_{10}$).

For the composition features $f_1$ and $f_2$, we first obtain the salient area of the images through salient detection [21] and then calculate the distance $d$ from the golden section and central points of the image according to the central position of the pixel coordinates in the significance area. After that, we divide the distance $d$ by the oblique length of the image to indicate the proximity to the composition. To numerically represent the positive correlation of the compositional feature values, we subtract this value by 1 and represent it as $f_1$ and $f_2$. The formula of the features is as follows:

$$f_{1,2} = 1 - \frac{d}{\sqrt{w^2 + h^2}}. \quad (11)$$

We obtain the composition feature line of the images by the model in Reference [8] and calculate $f_3$–$f_6$. First, we determine the composition feature lines according to the rule of thirds line, symmetrical line, diagonal line, and slash line. Then, we, respectively, calculate the distances between the two endpoints of each composition feature line and the two endpoints of the composition feature line, divide them by the long side of the image, and average them. Finally, we subtract this value with 1 to represent the compositional feature calculated values. To better distinguish the accuracy of the values, we set a threshold as 0.3, above which yields a positive value, or 0 otherwise. The calculation formula for $f_3$–$f_6$ is as follows:

$$f_{3-6} = 1 - \frac{d_1 + d_2}{2 \max(w, h)}. \quad (12)$$
Table 4. Composition Feature Values for Samples

| Composition feature | 11(a) | 11(b) | 11(c) | 11(d) | 11(e) | 11(f) | 11(g) | 11(h) | 11(i) | 11(j) |
|---------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Golden Section $f_1$ | 0.67  | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 0     |
| Center $f_2$         | 0     | 0.85  | 0     | 0     | 0     | 0.66  | 0.76  | 0     | 0.69  | 0.70  |
| Slant $f_3$          | 0     | 0     | 0.48  | 0     | 0.41  | 0     | 0     | 0.71  | 0     | 0     |
| Triangle $f_4$       | 0     | 0     | 0.44  | 0     | 0     | 0     | 0     | 0.71  | 0     | 0     |
| Guideline $f_5$      | 0     | 0     | 0     | 0.71  | 0     | 0     | 0     | 0     | 0     | 0     |
| Rule of Thirds $f_6$ | 0     | 0     | 0     | 0.48  | 0.84  | 0     | 0     | 0     | 0     | 0     |
| Symmetry $f_7$       | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 1     | 0     | 0     |
| Diagonal $f_8$       | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 1     | 0     | 0     |
| Frame $f_9$          | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 1     | 0     |
| Circle $f_{10}$      | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 1     |

We obtained features $f_7$–$f_9$ by calculating whether the intersection positions of the main composition lines are within the range of the images. If there is an intersection of the composition feature line on the image, then we will calculate the horizontal angle of the composition feature line. If more than a pair of horizontal angles are complementary, and the distances between the intersections of the composition feature lines are within $0.05\max(w, h)$, then $f_7 = 1$. If the sum of any three composition feature lines and the horizontal surface is $360^\circ$ and its intersection are within the image, then we will consider it as the triangle composition, $f_8 = 1$. If two composition feature lines are approximately parallel in the horizontal or vertical direction, and another composition feature line is almost vertical to them at the same time, then we will consider it as the frame composition, $f_9 = 1$. If we detect the image by the Hough circle and get the radius of the circle is longer than $0.1\min(w, h)$, then we will consider it as the circle composition, $f_{10} = 1$. Otherwise, all the above features $f_7$–$f_{10}$ are 0.

Figure 8 shows 10 sample images of different composition effects. Table 4 shows the composition feature values corresponding to these 10 images. From the feature values in Table 4, we can see that the images with one or more than one composition features have better composition effects.

5.4 Feature Fusion

The network structure in Figure 3 shows that the yellow part is the external attribute feature $f_{external}$ (4 for light features, 7 for color features, or 10 for composition features). We concat these features with the aesthetic high-level features extracted from the backbone network by the method mentioned in Section 4.2. The experiment results in Section 6 verify the method of the feature fusion.

6 EXPERIMENT

6.1 Training Details

We set the classification batch size to be 32, the regression batch size to be 64, and the learning rate to be 0.0001. We use Adam as the optimizer; betas are set as (0.98, 0.999); weight decay is set as 0.0001. If the accuracy rate of classification is not improved in two consecutive rounds, or the regression loss is not decreased in two consecutive rounds, the learning rate will multiply by 0.5. Our running environment is in MindSpore [26] 1.6.0 and Nvidia TITAN XPs. We divide the dataset into three sets; the ratio of training set and validation set and testing set is 8:1:1.
6.2 Analysis of Results

This article uses several indicators to measure training results:

**Mean square error (MSE)** represents the estimating errors of the difference between the predicted results and the true values; the formula is:

\[ MSE = \frac{1}{N} \sum_{i} (r_i - \hat{r}_i)^2. \]  
(13)

**SROCC (Spearman rankorder correlation coefficient, Spearman rank correlation coefficient)** represents the correlation between the predicted result and the true value and the formula is:

\[ SROCC = 1 - \frac{6 \sum_i (r_i - \hat{r}_i)^2}{N^3 - N}. \]  
(14)

The accuracy of the two-classification indicates whether the predicted score and the real score are consistent when the boundary is 5. The accuracy of two-classification indicates the most basic accuracy of classification; the formula is:

\[ ACCURACY = \frac{TP + TN}{P + N}. \]  
(15)

The positive and negative accuracy indicates whether the absolute value of the error is within 1 point between the predicted score and the real score. If the absolute value of the error is within 1 point, then the result can be considered accurate. The formula is:

\[ ACCURACY_{|error|\leq1} = \frac{N_{|error|\leq1}}{N}. \]  
(16)

We use EfficientNet-B0 as the baseline and conduct several experiments as follows: For feature fusion, Table 5 showed that external light, color, and composition attribute features could return to a more accurate assessment after fusion. For overall score regression, we used the soft loss
Aesthetic Attribute Assessment of Images Numerically

Table 5. Feature Fusion Experimental Results

| Methods                  | Attributes | MSE    | SROCC  | Acc   | Acc_{|error|≤1} |
|--------------------------|------------|--------|--------|-------|---------------|
| Baseline (MindSpore)     | Color      | 0.008660 | 0.6863 | 76.17% | 73.68%        |
|                          | Light      | 0.011266 | 0.6939 | 77.72% | 68.29%        |
|                          | Composition| 0.009491 | 0.6915 | 77.41% | 70.57%        |
| Basebone + Feature fusion (MindSpore) | Color      | 0.008307 | 0.7087 | 77.72% | 73.69%        |
|                          | Light      | 0.010675 | 0.7126 | 79.27% | 69.53%        |
|                          | Composition| 0.009285 | 0.7013 | 79.79% | 73.47%        |

Table 6. Soft Loss Experimental Results

| Methods                  | Attributes | MSE    | SROCC  | Acc   | Acc_{|error|≤1} |
|--------------------------|------------|--------|--------|-------|---------------|
| Baseline (MindSpore)     | Score      | 0.012940 | 0.8424 | 85.78% | 64.99%        |
| Baseline + Soft loss (MindSpore) | Score | 0.011315 | 0.8604 | 88.08% | 68.72%        |

Table 7. The Comparison Experiment for \( \lambda \)

| \( \lambda \) | MSE    | SROCC  | Acc   | Acc_{|error|≤1} |
|---------------|--------|--------|-------|---------------|
| 0             | 0.011408 | 0.8591 | 88.02% | 68.31%        |
| 0.1           | **0.011315** | **0.8604** | **88.08%** | **68.72%**    |
| 0.2           | 0.011318 | 0.8603 | 87.88% | 68.44%        |
| 0.3           | 0.011402 | 0.8595 | 87.81% | 68.58%        |
| 0.4           | 0.011437 | 0.8591 | 87.74% | 68.45%        |

mentioned in Section 4.2, and Table 6 showed that the model with soft loss had better performances in all indicators. We did a comparison experiment in 40 epoches of training. Table 7 shows that the presence of \( \lambda \) does improve the accuracy of the regression when \( \lambda = 0.1 \), so we set \( \lambda \) to 0.1.

Besides, we selected some typical samples and used this method to score the images, including the overall aesthetic score and the scores of the three aesthetic attributes. Among them, \( S \) represents the overall score, \( C \) represents the color score, \( L \) represents the light score, and \( CM \) represents the composition score, as shown in Figure 9.

7 SUMMARIZE

It is challenging to construct a new dataset in the image aesthetic quality assessment. Traditional datasets have the problem of few data and limited attribute labels. By mixing and filtering massive datasets and designing external attribute features, we get a new dataset called aesthetic mixed dataset with attributes (AMD-A) with a more reasonable distribution. Besides, we propose a model with multitasking network including one classification sub-network, three attribute sub-networks, and one regression sub-network. This is an innovative exploration of the training method for the numerical assessment of image aesthetics. Moreover, we design and use the external attribute feature fusion to improve the regressing of aesthetic attributes. According to the idea of the teacher-student network, we use the classification sub-network to guide the regression sub-network through high-level features by the soft loss. The experimental results proved that our model is more accurate than the traditional deep learning network regression model, and it improves the prediction of aesthetic attributes and overall scores.
Fig. 9. Test samples on the testing set with scores from low to high. S represents the overall score, C represents the color score, L represents the light score, and CM represents the composition score.
REFERENCES

[1] Javad Azimi, Ruofei Zhang, Yang Zhou, Vidhya Navalpakkam, Jianchang Mao, and Xiaoli Fern. 2012. The impact of visual appearance on user response in online display advertising. In 21st International Conference on World Wide Web. 457–458.

[2] Sean Bell, C. Lawrence Zitnick, Kativa Bala, and Ross Girshick. 2016. Inside-Outside Net: Detecting objects in context with skip pooling and recurrent neural networks. In IEEE Conference on Computer Vision and Pattern Recognition. 2874–2883.

[3] Anselm Brachmann and Christoph Redies. 2017. Computational and experimental approaches to visual aesthetics. Front. Computat. Neurosci. 11 (2017), 102.

[4] Kuang-Yu Chang, Kung-Hung Lu, and Chu-Song Chen. 2017. Aesthetic critiques generation for photos. In IEEE International Conference on Computer Vision. 3514–3523.

[5] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2006. Studying aesthetics in photographic images using a computational approach. In European Conference on Computer Vision. Springer, 288–301.

[6] Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2018. Aesthetic-driven image enhancement by adversarial learning. In 26th ACM International Conference on Multimedia. 870–878.

[7] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. 2020. Perceptual quality assessment of smartphone photography. In IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3677–3686.

[8] Qi Han, Kai Zhao, Jun Xu, and Ming-Ming Cheng. 2020. Deep Hough transform for semantic line detection. In European Conference on Computer Vision (ECCV). 249–265. DOI: https://doi.org/10.1007/978-3-030-58545-7_15

[9] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 2, 7 (2015).

[10] Gao Huang, Zhuang Liu, Laures Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In IEEE Conference on Computer Vision and Pattern Recognition. 4700–4708.

[11] Xin Jin, Le Wu, Xinghui Zhou, Geng Zhao, Xiaokun Zhang, Xiaodong Li, and Shiming Ge. 2018. Predicting aesthetic radar map using a hierarchical multi-task network. In Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Springer, 41–50.

[12] Xin Jin, Xinghui Zhou, Xiaokun Zhang, Hongbo Sun, Xiqiao Li, and Ruijun Liu. 2019. Incremental learning of multi-tasking networks for aesthetic radar map prediction. IEEE Access 7 (2019), 183647–183655.

[13] Magzhan Kairanbay, John See, and Lai-Kuan Wong. 2019. Beauty is in the eye of the beholder: Demographically oriented analysis of aesthetics in photographs. ACM Trans. Multim. Comput., Commun. Applic. 15, 2s (2019), 1–21.

[14] Chen Kang, Giuseppe Valenzise, and Frédéric Dufaux. 2020. EVA: An explainable visual aesthetics dataset. In Joint Workshop on Aesthetic and Technical Quality Assessment of Multimedia and Media Analytics for Societal Trends. 5–13.

[15] Yueying Kao, Chong Wang, and Kaiqi Huang. 2015. Visual aesthetic quality assessment with a regression model. In IEEE International Conference on Image Processing (ICIP). IEEE, 1583–1587.

[16] Yan Ke, Xiaoxu Tang, and Feng Jing. 2006. The design of high-level features for photo quality assessment. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06). IEEE, 419–426.

[17] Lu Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. In International Conference on Computer Vision. IEEE, 1784–1791.

[20] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In IEEE Conference on Computer Vision and Pattern Recognition. 2117–2125.

[21] Jiang-Jiang Liu, Qibin Hou, and Ming-Ming Cheng. 2020. Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton. IEEE Trans. Image Process. 29 (2020), 8652–8667.

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single shot multibox detector. In European Conference on Computer Vision. Springer, 21–37.

[23] Yiwen Luo and Xiaoou Tang. 2008. Photo and video quality evaluation: Focusing on the subject. In European Conference on Computer Vision. Springer, 386–399.

[24] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. In International Conference on Computer Vision. IEEE, 1784–1791.

[25] Shuang Ma, Jing Liu, and Chang Wen Chen. 2017. A-Lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In IEEE Conference on Computer Vision and Pattern Recognition. 4535–4544.
[27] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2408–2415.

[28] Masashi Nishiya, Takahiro Okabe, Imari Sato, and Yoichi Sato. 2011. Aesthetic quality classification of photographs based on color harmony. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 33–40.

[29] Pengfei Zhu, Peihua Li, Wangmeng Zuo, Qilong Wang, Banggu Wu, and Qinghua Hu. 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks. (2020), 11531–11539.

[30] Dongyu She, Yu-Kun Lai, Gaoxiong Yi, and Kun Xu. 2021. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 33–40.

[31] R. W. Sperry, M. S. Gazzaniga, and J. E. Bogen. 1969. Interhemispheric relationships: The neocortical commissures; syndromes of hemisphere disconnection. *Handb. Clin. Neurol.* 4 (1969), 145–153.

[32] Mingxiong Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 6105–6114.

[33] Hanghang Tong, Mingjing Li, Hong-Jiang Zhang, Jingrui He, and Changshui Zhang. 2004. Classification of digital photos taken by photographers or home users. In *Pacific-Rim Conference on Multimedia*. Springer, 198–205.

[34] Ranran Wang, Yin Zhang, Giancarlo Fortino, Qingxu Guan, Jiangchuan Liu, and Jeungeun Song. 2022. Software escalation prediction based on deep learning in the cognitive internet of vehicles. *IEEE Trans. Intell. Transport. Syst.* (2022).

[35] Ranran Wang, Yin Zhang, Limei Peng, Giancarlo Fortino, and Pin-Han Ho. 2022. Time-varying-aware network traffic prediction via deep learning in IIoT. *IEEE Trans. Industr. Inform.* (2022).

[36] Liming Xu, Xianhua Zeng, Weisheng Li, and Bochuan Zheng. 2022. MFGAN: Multi-modal feature-fusion for CT metal artifact reduction using GANs. *ACM Trans. Multim. Comput., Commun. Appl.* (2022).

[37] Yin Zhang, Yujie Li, Ranran Wang, Jianmin Lu, Xiao Ma, and Meikang Qiu. 2020. PSAC: Proactive sequence-aware content caching via deep learning at the network edge. *IEEE Trans. Netw. Sci. Eng.* 7, 4 (2020), 2145–2154.

[38] Peining Zhen, Shuqi Wang, Suming Zhang, Xiaotao Yan, Wei Wang, Zhigang Ji, and Hai-Bao Chen. 2022. Towards accurate oriented object detection in aerial images with adaptive multi-level feature fusion. *ACM Trans. Multim. Comput., Commun. Appl.* (2022).

[39] Quan Zhou, Xiaofu Wu, Suofei Zhang, Bin Kang, Zongyuan Ge, and Longin Jan Latecki. 2022. Contextual ensemble network for semantic segmentation. *Pattern Recog.* 122 (2022), 108290.

[40] Quan Zhou, Weining Yang, Guangwei Gao, Weihua Ou, Huimin Li, Jie Chen, and Longin Jan Latecki. 2019. Multiscale deep context convolutional neural networks for semantic segmentation. *World Wide Web* 22, 2 (2019), 555–570.

[41] Wei Zhou, Zhiwu Xia, Peng Dou, Tao Su, and Haifeng Hu. 2022. Double attention based on graph attention network for image multi-label classification. *ACM Trans. Multim. Comput., Commun. Appl.* (2022).

Received 4 December 2021; revised 24 May 2022; accepted 23 June 2022