

Received February 21, 2020, accepted March 18, 2020, date of publication March 27, 2020, date of current version April 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2983725

A Novel Feature Fusion Method for Computing Image Aesthetic Quality

XUEWEI LI^{1,2}, XUEMING LI^{1,3}, GANG ZHANG^{1,4}, AND XIANLIN ZHANG^{1,3}

¹Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing 100876, China

²School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

³School of Digital Media and Art Design, Beijing University of Posts and Telecommunications, Beijing 100876, China

⁴School of Armament Science and Technology, Space Engineering University, Beijing 101416, China

Corresponding author: Xuewei Li (lixuewei@bupt.edu.cn)

ABSTRACT Computationally, the aesthetic quality of an image means that the model automatically scores the aesthetic level of the image. However, there are many factors that determine beauty or ugliness for photographic photos. Therefore, extracting a variety of representative aesthetic features and fusing these features are still difficult tasks. In this paper, we design a two-stream network to calculate the aesthetic quality of the image. The upper stream of the network is an improved network with the SEResNet-50 and six skip connections added, which can improve the performance of the model without training to obtain deep convolutional neural network features. The lower stream of the network consists of the proposed algorithms for handcrafted extracting aesthetic features and multiple convolution layers to extract the aesthetic features. Finally, to fuse the features of the two-stream network without adding feature dimensions, a novel feature fusion layer is proposed. The results show that this novel feature fusion method can calculate results close to the artificial aesthetic evaluation.

INDEX TERMS Deep convolutional neural networks, feature fusion, handcrafted aesthetic features, image aesthetics quality assessment.


I. INTRODUCTION

Computational aesthetic quality of a photographic image means that a model can automatically calculate the aesthetic score of a photograph. If the photograph obtains a higher score, it demonstrates that the photograph is subjectively a “pretty” image. If the photograph earns a lower score, it indicates that the photograph is quite unattractive, even an “ugly” photograph. The aesthetic quality assessment has many appealing applications, such as managing personal photo albums, editing photos [1], retrieving images [2], and enhancing image quality [3]. As a result, computationally, the aesthetic quality of a photographic image has received increasing attention in the computer vision field [4]–[8].

The beauty or ugliness of a photograph often depends on many factors, such as whether the color of the image is vivid, whether it follows the rule of thirds, whether the subject matter is prominent, and whether the image has high resolution. Considering these factors, previous researchers have focused on handcrafted features [9]–[12]. However, the composition of photographs is very complicated, and handcrafted features

can only express one of the aesthetic features in the image composition. In the process of handcrafted feature extraction, there will be a few subjective preferences of the participants. Because of these limitations, most researchers have turned the focus of their research to autolearning features based on convolutional neural networks (CNN). By using the deep CNN to extract image high-level features, the subjective preference of extracting image features can be reduced [13]–[16]. Deep CNN features belong to the highly abstract features of the image, and it is difficult to explain which aesthetic features are extracted. Some researchers [22], [37] have attempted to merge handcrafted extracting features and CNN features through the concat layer, but the method increases the dimensions of the features, resulting in an increase in computational complexity.

Aesthetic features have the advantage of guiding features, but the extraction of aesthetic features is time-consuming and has a strong subjective bias. CNN features can be automatically extracted, and the process of the extraction is simple and convenient. However, the features extracted by CNN are highly abstracted, and as the depth of the network becomes deeper, the extracted features become more sparse. In order to fuse the advantages of the two features, we first designed

The associate editor coordinating the review of this manuscript and approving it for publication was Feng Shao .

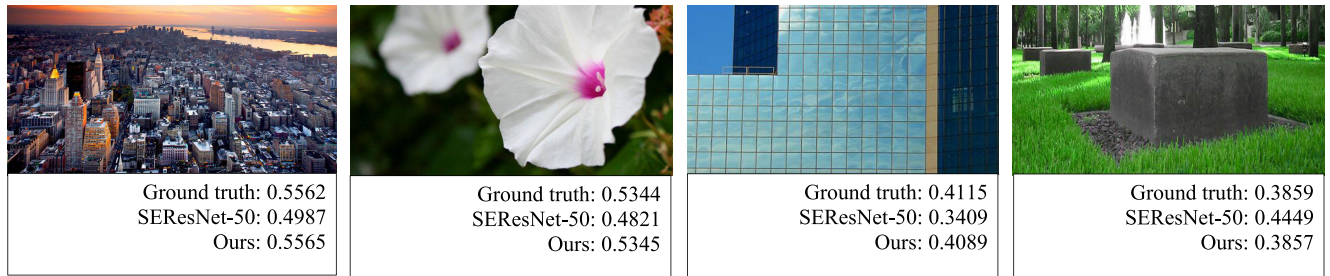


FIGURE 1. Some examples of predicting the aesthetic quality of images.

and extracted the aesthetic features, then merged the two features, and finally optimized the overall network structure. At present, there are two commonly used fusion algorithms: concat algorithm and eltwise algorithm. The concat algorithm will increase the spatial complexity of the algorithm, and more network parameters after fusion affect the training time of the network. The eltwise algorithm is easy to increase the computational complexity, and the fused features cannot distinguish the source of the features well. In order to overcome the problems of the two fusion algorithms, we propose and design an aesthetic feature fusion layer.

In this paper, we construct a two-stream network to assess image aesthetic quality. The upper stream of the network is an improved SEResNet-50 [41], which includes SEResNet-50 and six skip connections, to extract deep abstract features. The lower stream of the network consists of the proposed algorithms of handcrafted extracting aesthetic features and multiple convolution layers. A novel feature fusion layer is designed to fuse the features of the upper stream and lower stream. Some results of our method and SEResNet-50 [41] are shown in Fig. 1. Our contributions can be summarized as follows:

- A two-stream network is constructed. The upper stream is an improved SEResNet-50. The backbone network structure is optimized by adding six skip connections so that the performance of the backbone model can be improved without participating in training.
- The lower stream is a three-stage CNN structure that we designed, which is used to extract high-level aesthetic features.
- Five algorithms for handcrafted extracting aesthetic feature maps are proposed, which are used to extract feature maps of the brightness, color-harmony, rule of thirds, shallow depth of field, and motion blur of the image.
- A novel feature fusion layer is designed to fuse aesthetic features and CNN features.
- The experimental results show that the proposed feature fusion method can achieve more advanced results on the two public datasets.

The remainder of this paper is organized as follows. Section II briefly reviews previous works that are closely related to our method. A detailed description of our method is presented in Section III. The experiments are presented in

Section IV, and the conclusion of the paper is provided in Section V.

II. RELATED WORK

Aesthetic features are summarized by people according to the rules of photography [17], [18]. Researchers extract the aesthetic features of images based on their mission requirements and photography rules. Datta *et al.* [38] extracted 56 features for each image, which were divided into 9 categories: light and colorfulness, saturation and hue, the rule of thirds, a familiarity measure, wavelet-based texture, size and aspect ratios, region composition, low depth of field indicators and shape convexity. Ke *et al.* [39] found that high-quality and low-quality photos differ greatly in terms of simplicity, realism and basic techniques in the aesthetic quality classification task. Therefore, they extracted the edge space distribution of the image, the color distribution, the hue number and the blur as high-level features, and the contrast and brightness were extracted as low-level features of the image. Cohen-Or *et al.* [42] designed a conditional random field (CRF)-based color-harmony model to accomplish image aesthetic assessment tasks. However, the process of handcrafted features extraction is a very tedious and laborious task and requires a great deal of expert knowledge. And the traditional handcrafted aesthetic features are fixed and single values, which cannot express image aesthetics well.

Extracting features using CNN is performed automatically by the computer, and the high-level features of one entire image can be obtained without manual involvement and extensive expert knowledge. Tan *et al.* [19] introduced an aesthetic photo classifier with a deep and wide CNN, which can be applied to fine-grained aesthetic quality prediction. Kairanbay *et al.* [20] not only used deep CNN to extract aesthetic features but also used the global average pool to reduce the complexity of CNN. Xin *et al.* [21] proposed a method to take the global, local and scene perception information of an image into the deep CNN. The method extracted the abstract features by using the pretrained deep learning model and used the support vector machine to classify the extracted features. Even though the performance of the CNN-based method in the aesthetic quality classification task is significantly higher than that using handcrafted features, it is difficult to explain which kinds of aesthetic features the CNN extracted abstract features belong to.

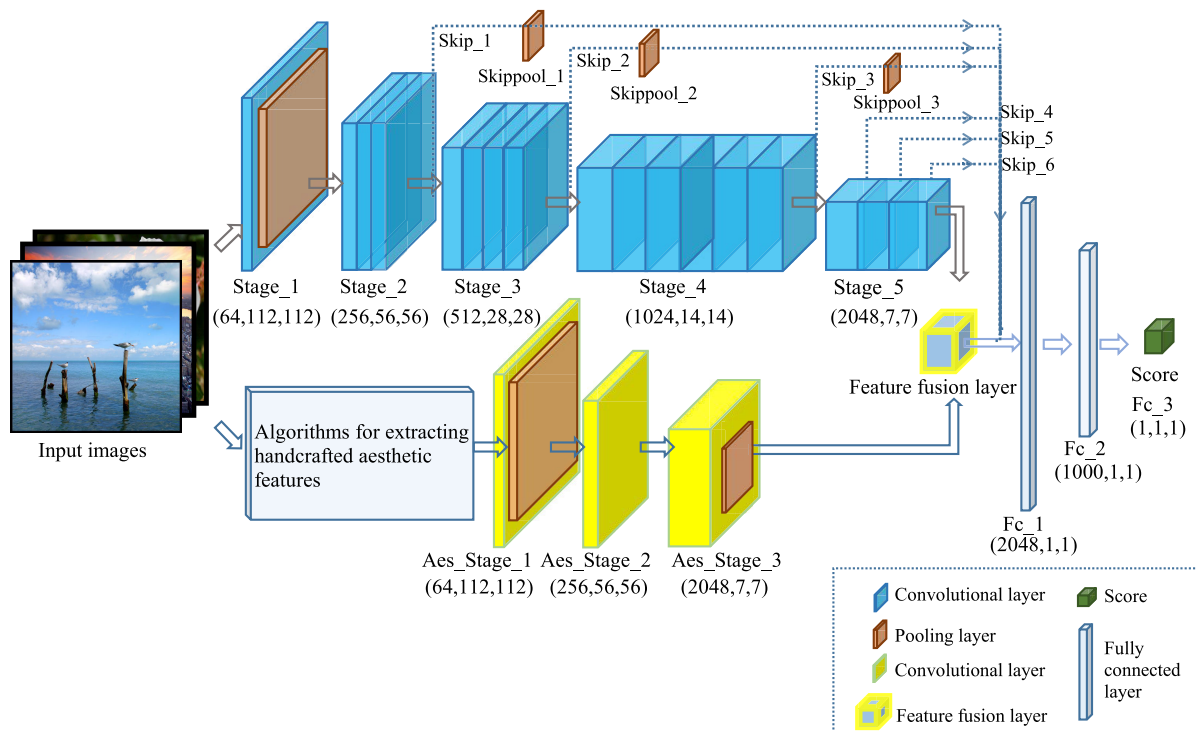


FIGURE 2. The network structure of the proposed method.

To overcome the drawbacks of handcrafted features and deep CNN features, many researchers have attempted to combine these two features. Kucer *et al.* [37] proposed the concept of early fusion and late fusion. Early fusion refers to the fusion of features before classification. Late fusion means that the classification decisions are separately made for multiple features, and then multiple decision results are combined. Late fusion is no longer the fusion of features but the fusion of classification results. Tian *et al.* [22] fused visual poetry features, text features and CNN features by early fusion and obtained very good aesthetic image classification results. There are two feature fusion layers in early fusion: the concat layer and the eltwise layer. The fusion of the concat layer leads to an increase in the feature dimension. The fusion of the eltwise layer does not ensure that the merged feature is a representation of the original image. For example, two features, “3” and “5”, are fused by the add method in the eltwise layer. The fused feature is “8” (3 + 5 = 8), but we cannot explain that feature “8” is a fusion from “3 add 5” or “4 add 4”.

In this paper, we design a new feature fusion layer. The dimensions of the fused features remain unchanged, and the fused features still ensure the representation of the image. The details of our fusion layer are described in Section III-C.

III. THE PROPOSED ALGORITHM

A. THE IMPROVED NETWORK STRUCTURE

The detailed network architecture we propose is displayed in Fig. 2. The network is primarily divided into the upper

stream, lower stream, a feature fusion layer and three fully connected layers. The input of the network is original RGB images, and the output of the last fully connected layer (Fc_3) is an aesthetic score. (n, w, h) under the Stage_1-Stage_5, the Aes_Stage_1-Aes_Stage_3 and Fc_1-Fc_3 indicate the number of channels and the width and height of the output features, respectively.

To extract the highly abstract global features of the image, the upper stream network, which consists of Stage_1-Stage_5 and Skip_1-Skip_6, is used. The structure of Stage_1-Stage_5 is the same as that of Conv_1-Conv_5_3 of SEResNet-50. Skip_1-Skip_6 are the six skip connections we added. There are two reasons for adding skip connections. First, it effectively solves the problem of a lack of image details in the global features extracted by Stage_5. The reason for the lack of image details in the features extracted by Stage_5 is that the framework of Stage_1-Stage_5 is a very deep network containing multiple convolution layers, which results in the obtained features having a very large receptive field. The larger the receptive field, the more abstract the feature description. To obtain as much image detail information as possible, features need to have small area receptive fields. In general, the features extracted by the shallow convolution of the network correspond to smaller receptive fields. Using skip connections can connect shallow convolutional features to deeper features so that the features finally extracted by the model can better capture image details. Second, to solve the problem of using pretrained models on new tasks that require retraining or

fine-tuning, the pretrained model of SEResNet-50 is actually obtained in the classification task on ImageNet [13], and it does not perform the image aesthetic quality assessment task well. The retraining of the SEResNet-50 model also lacks sufficient annotated datasets and takes considerable time to obtain the model to converge. Therefore, by adding six skip connections, the network can effectively increase the performance of the model without retraining. Skipool_1-Skipool_3 are the three average pooling layers. The kernel size of these pooling layers are 3, 7 and 5, respectively.

To obtain the aesthetic features of the image, we design the lower stream structure of the network. The lower stream consists of two parts: the algorithms of handcrafted aesthetic feature extraction and Aes_Stage_1-Aes_Stage_3. The handcrafted aesthetic feature algorithms are described in detail in Section III-B. The aesthetic features of input images are extracted by handcrafted algorithms, and these aesthetic features are sent to Aes_Stage_1. Aes_Stage_1, Aes_Stage_2 and Aes_Stage_3 are used to extract deep aesthetic features. The extraction of deep features makes up for the lack of handcrafted algorithms. The structure of Aes_Stage_1-Aes_Stage_2 comes from Conv_1-Conv_2_1 of SEResNet-50. The input of Aes_Stage_1 is the handcrafted aesthetic features, and the feature size is $224 \times 224 \times 3$. The Aes_Stage_1 is consisted of a 7×7 convolution layer and a 3×3 pooling layer. The Aes_Stage_3 is composed of a 3×3 convolution layer and a 4×4 pooling layer. The reason for designing Aes_Stage_3 is that it can change the number of channels of the feature and increase the channel from 256 to 2,048. To be consistent with the number of global features extracted from the upper stream of the network, the number of channels is set to 2,048.

At the end of the network, a novel feature fusion layer is used to fuse the aesthetic features and the global deep CNN features. The specific fusion method is described in Section III-C. Then, the fused features are sent to three fully connected layers, and finally, the aesthetic score of an image can be calculated.

B. THE ALGORITHMS OF HANDCRAFTED AESTHETIC FEATURE EXTRACTION

The photography books [17], [18] are references for the techniques professional photographers often use to improve the quality of their work. By carefully comparing the aesthetics with attributes database (AADB) [31] and aesthetic visual analysis (AVA) [30] datasets, five aesthetic features are finally selected. Therefore, the five algorithms of handcrafted aesthetic features extraction are designed and used to extract the aesthetic features of brightness, color-harmony, rule of thirds, shallow depth of field and motion blur. These five aesthetic features are carefully selected, but the number is limited. Because our main goal is to study the fusion of any aesthetic features to effectively improve the performance of the model, limited aesthetic features will not affect our research.

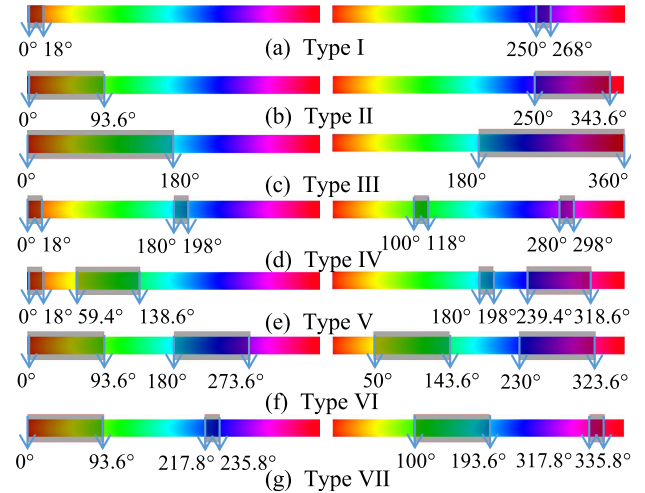


FIGURE 3. Seven color-harmony types.

TABLE 1. The width of each gray rectangle.

Type	Width_first	Width_second	Distance
Type I	18°	-	-
Type II	93.6°	-	-
Type III	180°	-	-
Type IV	18°	18°	180°
Type V	18°	79.2°	59.4°
Type VI	93.6°	93.6°	180°
Type VII	93.6°	18°	217.8°

1) COLOR-HARMONY

As one of the important features, color provides plenty of useful information to represent images. A set of harmonious colors can be used to create visually pleasing feelings. Cohen-Or *et al.* [42] described eight color-harmony types defined over the hue channel of the HSV color wheel. Inspired by Olga Sorkine’s work, we define seven harmonic types over the hue channel of the HSV color bar. The eighth type mentioned in Olga Sorkine’s work is a grayscale image, but grayscale images are not discussed in this paper.

As shown in Fig. 3, seven harmonic types are displayed: type I, type II, type III, type IV, type V, type VI and type VII. Each type has two color bars (left and right). There are one or two gray rectangles on each color bar, and each gray rectangle can slide freely left and right. If the hue value of an image is within the gray rectangular range, the color of the image is considered to be harmonious, and vice versa. The width of each gray rectangle is listed in Table 1, where “Width_first” denotes the width of the first rectangle, “Width_second” denotes the width of the second rectangle, and “Distance” is the distance between the left lines of the two rectangles.

The entire process of calculating an image color-harmony feature map is described in Algorithm 1. Here, we need to explain the two color spaces first: the RGB color space and the HSV color space. The I_R , I_G and I_B represent the red, green and blue values, respectively. The I_H , I_S and I_V represent the hue, saturation and brightness values, respectively.

Note that, the I_R , I_G , I_B , I_H , I_S and I_V are two-dimensional matrices, and the width and height of each two-dimensional matrix are w and h , respectively.

Algorithm 1 The Proposed Method for Calculating an Image Color-Harmony Feature Map

Input: The RGB data of an image

Output: The color-harmony feature map of the image

STEP1: The RGB data of the image need to be converted to the HSV color space.

STEP2: Calculate the histogram of I_H .

repeat

STEP3: Calculate the distance between each angle of a color-harmony type and histogram of I_H . The position of each gray rectangle is the angle of that type. The angle ranges from 0° to 360° . The angle of minimum distance is considered the best angle of this type.

until The best angles and minimum distances are calculated for all seven types.

STEP4: Select the minimum distance from the seven distances. The color-harmony type corresponding to the minimum distance is the optimal type.

STEP5: Retain the I_H values within the range of the optimal type, and retain the I_S and I_V values at the same pixel position. Values at other positions are replaced by image gray values.

STEP6: Convert new I_H , I_S and I_V values to I_R , I_G and I_B values, and then the color-harmony feature map can be obtained.

Examples of color-harmony and color-disharmony are shown in Fig. 4. It is easy to find that the color-harmony feature map can show more colors than the color-disharmony feature map.

2) RULE OF THIRDS

The “rule of thirds” is one of the most basic concepts in art, and it refers to an attractive and artistic method of composition. The rule of thirds states that when the entire image is divided into one-third, the most important elements in the image of the combination should be placed around these four lines. The rule of thirds helps people quickly capture a clear picture of the subject when shooting complex scenes. It is worthy noting that only the subject follows the rule of thirds in the entire image, and the background does not need to satisfy this rule. It can be clearly seen from Fig. 5 that Fig. 5(a) satisfies the rule of thirds better than Fig. 5(b). The calculation details of the rule of thirds feature maps (f_{rot}) can be defined as:

$$f_{rot} = \{I_R(x, y), I_G(x, y), I_B(x, y) \mid \times \frac{1w}{3} \leq x \leq \frac{2w}{3}, \frac{1h}{3} \leq y \leq \frac{2h}{3}\}, \quad (1)$$

where x and y are the positions in the image.

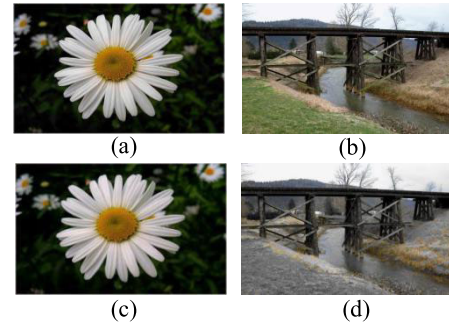


FIGURE 4. Examples of the color-harmony and color-disharmony. (a) is an image with color-harmony. (b) is an image with color-disharmony. (c) and (d) are the feature maps of (a) and (b), respectively.

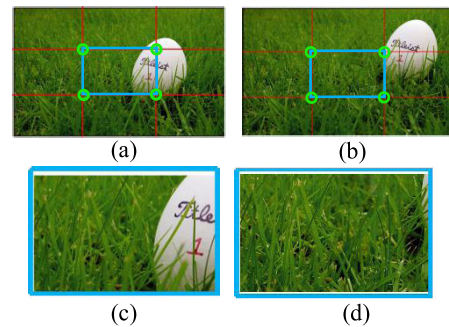


FIGURE 5. Examples of the rule of thirds. (a) and (b) are images that have the same content but different object positions. (c) and (d) are the feature maps of (a) and (b), respectively.

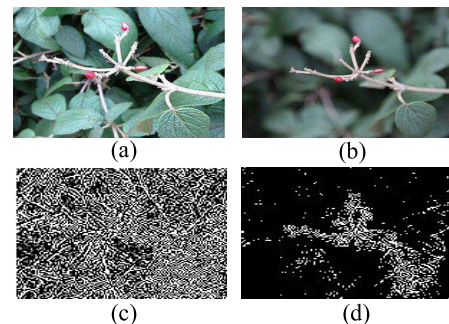


FIGURE 6. Examples of a shallow depth of field feature map. (a) is an image without shallow depth of field. (b) is an image with a shallow depth of field. (c) and (d) are the feature maps of (a) and (b), respectively.

3) SHALLOW DEPTH OF FIELD

The “shallow depth of field” indicator is used to measure whether the image has a shallow depth of field. The transition of image content from sharp focus to soft focus is called the depth of field. The shallow depth of field can help to quickly attract the attention of an audience to the theme of the image. Examples of shallow depth of field feature map extraction are shown in Fig. 6.

To obtain the feature maps of shallow depth of field, the following steps need to be performed:

1) To obtain the transformed high-frequency image in both directions $I_{HH}(x, y)$, a Haar wavelet transform [23] needs to be performed on the image.

-1	0
0	1

0	-1
1	0

1	1	1
1	-8	1
1	1	1

(a) Roberts operator (b) Laplacian operator

FIGURE 7. The value of two operators.

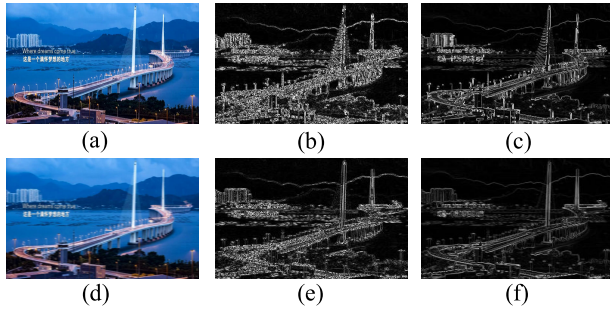


FIGURE 8. Examples of detecting motion blur. (a) is an image without motion blur. (d) is an image with motion blur. (b) and (e) are the results of Laplacian operator edge detection on (a) and (d), respectively. (c) and (f) are the results of Roberts operator edge detection on images (a) and (d), respectively.

2) Calculate a threshold to filter the high-frequency image:

$$T = |\text{round}(\frac{I_{max} - (I_{max} - I_{min})}{2}) - 5| \quad (2)$$

where I_{max} and I_{min} can be denoted as:

$$I_{max} = \max_y(\max_x(I_{HH}(x, y))), \quad (3)$$

$$I_{min} = \min_y(\min_x(I_{HH}(x, y))), \quad (4)$$

where \max_x and \max_y are the maximum values in the x and y directions, respectively, and \min_x and \min_y are the minimum values in the x and y directions, respectively. The round is a rounding function.

3) If the pixel value of $I_{HH}(x, y)$ is larger than the threshold (T), set it to 1. If the pixel value of $I_{HH}(x, y)$ is smaller than the threshold (T), set it to 0.

4) MOTION BLUR

“Motion blur” is an indicator that can determine whether image blur has been caused by motion. If the image has motion blur, the image quality is generally low. Here, we do not discuss special images that are specifically shot with motion blur. It is interesting to note that images with motion blur also have edge blur problems. Inspired by this phenomenon, we attempt to use the Roberts [26] and Laplacian [24], [25] operators to perform edge detection. Because these two operators have an edge sensitivity in common, they can perform edge detection tasks well. The specific values of the two operators in the experiment are shown in Fig. 7.

As shown in Fig. 8, the images that do not have motion blur are more likely to detect more edges than the images that have motion blur. From the experimental results of the edge detection by two operators, we can conclude that the results

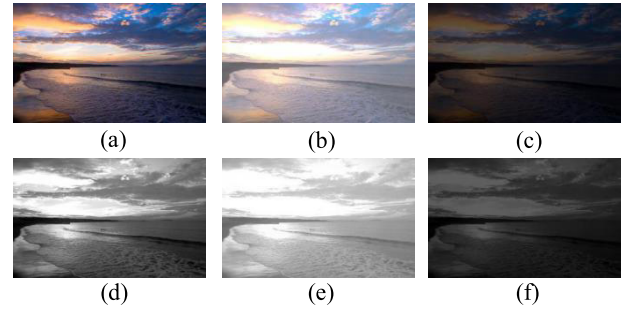


FIGURE 9. Some examples with different brightness. (a) A suitable brightness image. (b) An excessive brightness image. (c) A lack of brightness image. (d), (e) and (f) are the brightness feature maps of (a), (b) and (c), respectively.

obtained by using the Laplacian operator are rich in image texture information, while the results of the Roberts operator detection have less. Therefore, the Laplacian operator is selected as the operator for extracting the feature maps of the image motion blur.

5) BRIGHTNESS

The “brightness” feature is a rule that measures whether an image has good or interesting lighting. Some examples with different brightness are shown in Fig. 9. If there is too much light in the image, the photo is brighter. If there is not enough light, the photo is too dark and often unattractive. Therefore, high-quality and low-quality photos are well distinguished by brightness. Please note that there are exceptions to any “rule of thumb”. In some cases, an overexposed or underexposed photo may produce a very novel and beautiful photo. In this paper, we do not discuss special cases such as this. To obtain the brightness feature maps, images should be converted from the RGB color space to HSV color space. Since I_V represents the brightness value of the image, we directly use the value of I_V to represent the image brightness feature map.

C. THE PROPOSED FEATURE FUSION LAYER

The fusion layers are provided in many deep learning frameworks, such as Caffe, TensorFlow and PyTorch. The concat layer and eltwise layer are provided by Caffe. Both of these layers use early fusion. The role of the concat layer is to splice two or more feature maps in the channel or number dimensions. If splicing is performed on the channel dimensions, the spliced features increase in the dimensions of the channel. If splicing is performed on the number dimensions, then the spliced features increase in the number dimensions. The eltwise layer completes fusion by performing a point-to-point operation on the two feature maps and generally performs a point-to-point summing operation. It is not difficult to find that the point-to-point operation will not increase the dimensionality of the features and increase the number of calculations. Even then, the eltwise layer has a disadvantage: it is impossible to distinguish whether the fused feature is an aesthetic feature or a global feature. As shown in Fig. 10, the

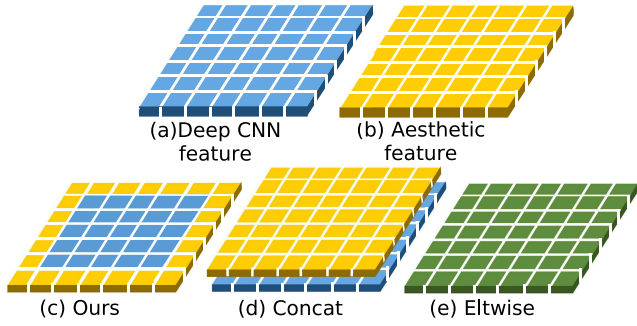


FIGURE 10. Examples of three fused methods .

proposed feature fusion layer does not increase the number of computations and easily distinguishes the source of the merged feature.

Our proposed fusion method replaces features at the edges of the deep CNN feature map with features of the edge of the aesthetic feature map. There are two reasons for this. First, by analyzing the deep CNN feature maps, we find that the features are very sparse, and most edge position feature values of the feature maps are zero. Second, we find that the receptive field at the edge feature is also at the edge of the original image. After the edge features of the deep feature map are replaced by aesthetic features, it does not affect the performance of the original global deep CNN features.

The training process for our proposed feature fusion layer is illustrated in the Fig. 11, where FP_{cnn} , FP_{aes} and FP_{fusion} represent the deep CNN, aesthetics and fused feature maps, respectively, and the matrices of these feature maps are listed as (5), (6) and (7). Fc_1 represents the first fully connected layer of our network. From left to right is the process of forward propagation, and from right to left is the process of backward propagation. In the forward propagation processing, the features of FP_{cnn} and FP_{aes} are fused into a new feature FP_{fusion} , which is then fed into Fc_1 . In backward propagation processing, after the parameters of the Fc_1 layer are all derived, they are propagated to the feature fusion layer, and the feature fusion layer passes the derivative results to FP_{cnn} and FP_{aes} .

$$FP_{cnn} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & \dots & c_{17} \\ c_{21} & c_{22} & c_{23} & \dots & c_{27} \\ c_{31} & c_{32} & c_{33} & \dots & c_{37} \\ \dots & \dots & \dots & \dots & \dots \\ c_{71} & c_{72} & c_{73} & \dots & c_{77} \end{bmatrix} \quad (5)$$

$$FP_{aes} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{17} \\ a_{21} & a_{22} & a_{23} & \dots & a_{27} \\ a_{31} & a_{32} & a_{33} & \dots & a_{37} \\ \dots & \dots & \dots & \dots & \dots \\ a_{71} & a_{72} & a_{73} & \dots & a_{77} \end{bmatrix} \quad (6)$$

$$FP_{fusion} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{17} \\ a_{21} & c_{22} & c_{23} & \dots & a_{27} \\ a_{31} & c_{32} & c_{33} & \dots & a_{37} \\ \dots & \dots & \dots & \dots & \dots \\ a_{71} & a_{72} & a_{73} & \dots & a_{77} \end{bmatrix} \quad (7)$$

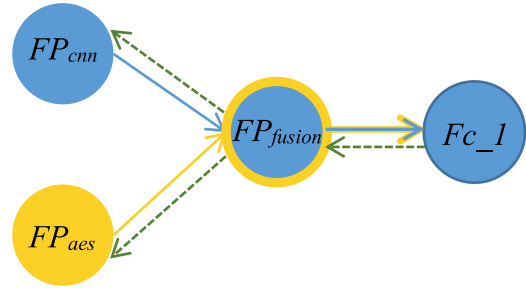


FIGURE 11. The training process of our proposed feature fusion layer.

In order to better compare the computing performance of the three fusion methods, we analyze their computational complexity in the next section. As we all know, computational complexity includes space complexity and time complexity. As is shown in (5), (6) and (7), the aesthetic features and CNN features are one 7×7 matrix, respectively. The proposed feature fusion method requires 24 steps to obtain a 7×7 fused feature matrix. The eltwise algorithm requires 49 steps to obtain a 7×7 fused feature matrix. The time complexity of the eltwise algorithm is 49/24 times that of our algorithm. Although the concat algorithm does not need to perform mathematical calculations, the features after fusion are two 7×7 matrices. The space complexity of the concat algorithm is twice that of our proposed algorithm.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The performance of the feature fusion method proposed in this paper was verified by the experiments in this section. Additionally, six objective evaluation indicators were used to judge the experimental results. Compared with other advanced methods, the experimental results of our model achieved the most advanced performance. We cover these experiments in detail in the rest of this section. To control the uniqueness of the variables in the experiment, all experiments were trained under the same conditions. For example, all models are trained on Caffe. The batch size of the training data during the training is 14, and the batch size of the validating data during the training is 10. The momentum is 0.9. The weight decay value is 0.0005. The optimizer uses stochastic gradient descent. In addition, all models were trained and tested on the PCs with Intel i7-8700k Intel CPUs and NVIDIA 1080Ti GPUs.

A. DATASETS

Data widely used in the field of aesthetic quality assessment are the aesthetic visual analysis dataset (AVA) [30], aesthetics with attributes database (AADB) [31], CUHK [39], CUKHPQ [28] and PhotoNet [38] datasets. These data are often collected by crawling from online communities such as www.photo.net and www.dpchallenge.com. The data producers ask professional or amateur photographers to mark the images, which usually represent the aesthetic judgment of professional or amateur photographers. We believe that such

annotations are highly significant because they capture the way amateurs and professionals understand the aesthetics of images. It is based on such artificial evaluation criteria that different methods can be used to fairly assess aesthetics of the natural image, and the aesthetic scores of the automatic machine judgment are quantitatively compared with the manual score.

1) AVA DATASET

The AVA dataset is a large image aesthetic database. The AVA is a collection of images derived from www.dpchallenge.com. This dataset contains more than 250,000 images and has a rich variety of image annotations. The annotations contained in the AVA dataset can be broadly divided into three types: aesthetic annotations, segment annotations and photostyle annotations. In the aesthetic annotations, each image was evaluated for aesthetic assessment by approximately 200 individuals. Each person voted for an image based on aesthetic criteria from one to ten points. In the experiment, all aesthetic scores were normalized. We randomly selected 30,000 images from 250,000 images for experimental training and validation data, of which 28,000 were used as training datasets, and 2,000 were used as validation datasets. To compare experimental results impartially with other methods, we obtained 20,000 images as test data in the same way as in the literature AVA [30], AADB [31] and Rapid [16].

2) AADB DATASET

The AADB dataset not only contains many photos taken by professional photographers but also contains photos taken by many amateur photographers, so the AADB dataset is more balanced in the distribution of photos than the AVA dataset. The AADB dataset also has more levels of attribute tags than the AVA dataset. To collect a large variety of photographic images, the images in the AADB dataset were downloaded from the Flickr website. Each image has an aesthetic score of 0 to 1 point, which is provided by five different evaluators. Since the AADB dataset contains a total of 10,000 images, in our experiment, 8,500 images were used as the training dataset, 500 images were used as the validation dataset, and the remaining 1,000 images were used as the testing dataset. We split the whole AADB dataset in the same way as in [31].

B. EVALUATION METRICS

The reliability of the proposed method is confirmed by the results of six commonly used objective evaluation metrics. The six evaluation metrics are the mean square error (MSE), mean absolute error (MAE), median absolute error (MED), Pearson correlation coefficient ($P_{-}\rho$), Spearman correlation coefficient ($Sp_{-}\rho$) and classification accuracy (Acc).

The MSE is a measure of the degree of difference between the predicted value of the model and the artificial score. The formula for MSE can be written as

$$MSE_{(g,p)} = \frac{1}{N} \sum_{i=1}^N (g_i - p_i)^2 \quad (8)$$

where g is the score of the artificial aesthetic assessment of the image, and p is the aesthetic score of the model prediction. N is the total number of test datasets. The value of N is 20,000 in the AVA dataset and 1,000 in the AADB dataset. i represents the i -th image in the testing data. The value of i ranges from 1 to N .

The MAE has the same effect as the MSE and is the average of the absolute values of the deviations between the model predictor and the artificial score. The MAE can avoid the problem that the errors cancel each other out, so it can accurately reflect the actual prediction error. The formula for MAE is shown below:

$$MAE_{(g,p)} = \frac{1}{N} \sum_{i=1}^N |g_i - p_i|. \quad (9)$$

The MED is the absolute value of the difference between the model prediction value and the artificial score of all test samples. All the differences are arranged, and the median value is taken as the median absolute error. The MED is suitable for datasets with outliers. The MED can effectively avoid the influence of outliers on the experimental results. The MED is formulated as

$$MED_{(g,p)} = med(|g_1 - p_1|, |g_2 - p_2|, \dots, |g_i - p_i|), \quad (10)$$

where med represents the median of the computed array.

$P_{-}\rho$ is used to measure the linear correlation between two variables, whose values are between -1 and 1. The $P_{-}\rho$ can be expressed as

$$P_{-}\rho_{(g,p)} = \frac{Cov(g,p)}{\sqrt{D(g)}\sqrt{D(p)}}, \quad (11)$$

where $Cov(g,p)$ can be written as

$$Cov(g,p) = \frac{\sum_{i=1}^N (g_i - \bar{g})(p_i - \bar{p})}{N}, \quad (12)$$

where \bar{g} and \bar{p} are the mean of the ground truth and prediction, respectively. Then, $\sqrt{D(g)}$ and $\sqrt{D(p)}$ can be defined as:

$$\sqrt{D(g)} = \sqrt{\frac{1}{N} \sum_{i=1}^N (g_i - \bar{g})^2}, \quad (13)$$

$$\sqrt{D(p)} = \sqrt{\frac{1}{N} \sum_{i=1}^N (p_i - \bar{p})^2}. \quad (14)$$

Finally, $P_{-}\rho$ can be written as:

$$P_{-}\rho_{(g,p)} = \frac{\sum_{i=1}^N (g_i - \bar{g})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^N (g_i - \bar{g})^2} \sqrt{\sum_{i=1}^N (p_i - \bar{p})^2}} \quad (15)$$

$Sp_{-}\rho$ is defined as the Pearson correlation coefficient between the gradation variables. The formula for calculating $Sp_{-}\rho$ can be represented as:

$$Sp_{-}\rho_{(Rg,Rp)} = \frac{\sum_{i=1}^N (Rg_i - \bar{Rg})(Rp_i - \bar{Rp})}{\sqrt{\sum_{i=1}^N (Rg_i - \bar{Rg})^2} \sqrt{\sum_{i=1}^N (Rp_i - \bar{Rp})^2}}, \quad (16)$$

TABLE 2. The performance of different backbone CNN models on the AADB dataset.

Models	MSE ↓	MAE ↓	MED ↓	P _{-ρ} ↑	Sp _{-ρ} ↑	Acc(%) ↑
AlexNet	0.0389	0.1611	0.1375	0.5676	0.5637	61.2000
GoogleNet	0.0270	0.1331	0.1150	0.5991	0.5976	74.8000
ResNet-50	0.0313	0.1408	0.1199	0.5673	0.5519	71.2000
SEResNet-50	0.0243	0.1236	0.1068	0.6627	0.6552	75.1000

TABLE 3. The performance of different backbone CNN models on the AVA dataset.

Models	MSE ↓	MAE ↓	MED ↓	P _{-ρ} ↑	Sp _{-ρ} ↑	Acc(%) ↑
AlexNet	0.0082	0.0723	0.0640	0.3200	0.3154	66.0600
GoogleNet	0.0066	0.0657	0.0562	0.2806	0.2724	73.5250
ResNet-50	0.0053	0.0577	0.0481	0.5183	0.5134	77.0850
SEResNet-50	0.0045	0.0528	0.0440	0.6073	0.5995	80.1550

where Rg and Rp are the grades of the manual score and the model prediction score, respectively, and $\bar{R}g$ and $\bar{R}p$ represent the average of the manual score and the model prediction grade.

By simply thresholding the predicted aesthetic score, we can obtain a binary classification accuracy. When an image obtains a score higher than 5, it is a high-quality image. Aesthetic scores below 5 are considered to be low-quality images. Assume that TH_p is the correct number of predicted high-quality image, and TL_p is the correct number of predicted low-quality image. Thus, Acc can be defined as:

$$Acc = \frac{TH_p + TL_p}{N}. \quad (17)$$

C. CHOOSING A BACKBONE CNN NETWORK

Since the CNN model solves various problems in the field of computer vision and achieves good results, it becomes easier to understand using CNN to extract image features. As seen from a recent review by Deng *et al.* [8], many ideas for CNN model design are derived from AlexNet [13], which achieved the most advanced results for the first time in the ImageNet competition. We fine-tuned four popular CNN models (such as AlexNet [13], GoogleNet [14], ResNet50 [40], SENet [41]) on the AADB and AVA datasets, which showed good results in both classification and regression tasks. To verify the performance of the four network prediction aesthetic scores, we did not significantly modify the structure of the four benchmark networks during the experiment but set the output of the last fully connected layer to 1 and used this output as an aesthetic score for automatic prediction of the model.

Table 2 and Table 3 show the results of the four models on the AADB dataset and the AVA dataset. By numerical comparison, it can be found that the best performing of the four models is SEResNet-50. Therefore, in subsequent experiments, SEResNet-50 was selected as the backbone network. After the evaluation indicators in the table of this section, an upward or downward arrow is drawn. The upward arrow (“↑”) indicates that the higher the value of the indicator, the better. The downward arrow (“↓”) implies that the lower the

TABLE 4. Methods for comparing three fusion brightness features on the AADB dataset.

Fusion type	MSE ↓	MAE ↓	MED ↓	P _{-ρ} ↑	Sp _{-ρ} ↑	Acc(%) ↑
No	0.0243	0.1236	0.1068	0.6627	0.6552	75.1000
Concat	0.0234	0.1218	0.0981	0.6654	0.6567	76.9000
Eltwise	0.0239	0.1227	0.1044	0.6645	0.6616	75.9000
Ours	0.0226	0.1186	0.0975	0.6785	0.6742	76.6000

TABLE 5. Methods for comparing three fusion brightness features on the AVA dataset.

Fusion type	MSE ↓	MAE ↓	MED ↓	P _{-ρ} ↑	Sp _{-ρ} ↑	Acc(%) ↑
No	0.0045	0.0528	0.0440	0.6073	0.5995	80.1550
Concat	0.0040	0.0499	0.0413	0.6519	0.6433	81.8500
Eltwise	0.0041	0.0501	0.0417	0.6496	0.6409	81.6800
Ours	0.0040	0.0498	0.0413	0.6524	0.6461	82.0800

TABLE 6. The fusion results of aesthetic feature on the AADB dataset.

Feature type	f	MSE ↓	MAE ↓	MED ↓	P _{-ρ} ↑	Sp _{-ρ} ↑	Acc(%) ↑
No	-	0.0243	0.1236	0.1068	0.6627	0.6552	75.1000
Brightness	c	0.0234	0.1218	0.0981	0.6654	0.6567	76.9000
	e	0.0239	0.1227	0.1044	0.6645	0.6616	75.9000
	o	0.0226	0.1186	0.0975	0.6785	0.6742	76.6000
Color-H	c	0.0242	0.1225	0.1006	0.6606	0.6508	76.2000
	e	0.0248	0.1243	0.1051	0.6475	0.6441	75.1000
	o	0.0236	0.1209	0.0998	0.6677	0.6600	76.3000
Rule of thirds	c	0.0257	0.1265	0.1063	0.6476	0.6365	74.7000
	e	0.0290	0.1332	0.1097	0.6007	0.5996	72.5000
	o	0.0240	0.1220	0.1026	0.6615	0.6497	76.7000
S-dof	c	0.0242	0.1221	0.1000	0.6605	0.6508	76.6000
	e	0.0240	0.1229	0.1055	0.6609	0.6565	76.2000
	o	0.0235	0.1219	0.1013	0.6693	0.6590	77.1000
Motion blur	c	0.0242	0.1226	0.1004	0.6630	0.6548	77.5000
	e	0.0259	0.1282	0.1087	0.6598	0.6570	75.4000
	o	0.0238	0.1223	0.1000	0.6680	0.6558	76.5000

value of the indicator, the better. We show the best results in bold.

D. IMPLEMENTATION AND ANALYSIS OF THREE FEATURE FUSION METHODS

As shown in Table 4 and Table 5, we first attempt to fuse the brightness feature of the image with the global features extracted by SEResNet-50. Table 4 and Table 5 give four experimental results: “No” indicates that the model does not fuse any aesthetic features, “Concat” indicates fusion using the concat layer, “Eltwise” denotes fusion using the eltwise layer, and “Ours” is fusion using our proposed method. As shown by the bold results in the table, our proposed fusion method performed best in the aesthetic evaluation task.

According to the experimental results of the brightness feature, we can find that the proposed fusion algorithm has a significant improvement. Therefore, in order to fully verify the effectiveness of the proposed fusion layer of aesthetic features, we also performed ablation experiments on the other four aesthetic features. The results of the fusion

TABLE 7. The fusion results of aesthetic feature on the AVA dataset.

Feature type	f	MSE ↓	MAE ↓	MED ↓	$P_{-}\rho$ ↑	$Sp_{-}\rho$ ↑	Acc(%) ↑
No	-	0.0045	0.0528	0.0440	0.6073	0.5995	80.1550
Brightness	c	0.0040	0.0499	0.0413	0.6519	0.6433	81.8500
	e	0.0041	0.0501	0.0417	0.6496	0.6409	81.6800
	o	0.0040	0.0498	0.0413	0.6524	0.6461	82.0800
Color-H	c	0.0041	0.0504	0.0418	0.6452	0.6358	81.3450
	e	0.0043	0.0516	0.0423	0.6215	0.6154	80.5600
	o	0.0040	0.0498	0.0413	0.6524	0.6463	82.1150
Rule of thirds	c	0.0044	0.0519	0.0427	0.6126	0.6063	80.8750
	e	0.0056	0.0583	0.0476	0.5551	0.5542	76.3750
	o	0.0040	0.0498	0.0414	0.6527	0.6466	81.9750
S-dof	c	0.0044	0.0528	0.0439	0.6321	0.6211	79.8800
	e	0.0042	0.0512	0.0425	0.6417	0.6322	80.5600
	o	0.0041	0.0500	0.0416	0.6483	0.6416	81.8600
Motion blur	c	0.0042	0.0513	0.0426	0.6449	0.6361	80.2250
	e	0.0042	0.0515	0.0428	0.6435	0.6350	80.0950
	o	0.0040	0.0498	0.0414	0.6522	0.6459	82.0750

TABLE 8. The fusion results of aesthetic feature on the mCUHK dataset.

Feature type	f	MSE ↓	MAE ↓	MED ↓	$P_{-}\rho$ ↑	$Sp_{-}\rho$ ↑	Acc(%) ↑
No	-	0.2024	0.4454	0.4457	0.6795	0.7065	80.4000
Brightness	c	0.1992	0.4422	0.4422	0.7151	0.7321	82.6000
	e	0.2011	0.4444	0.4421	0.7120	0.7296	80.8000
	o	0.1989	0.4413	0.4400	0.7197	0.7301	84.2000
Color-H	c	0.2005	0.4439	0.4446	0.7052	0.7217	80.2000
	e	0.2015	0.4450	0.4431	0.6817	0.6997	82.8000
	o	0.1997	0.4414	0.4395	0.7173	0.7277	83.2000
Rule of thirds	c	0.2013	0.4445	0.4440	0.6816	0.6971	79.7000
	e	0.2017	0.4452	0.4421	0.6087	0.6269	81.8000
	o	0.2011	0.4400	0.4327	0.7149	0.7254	83.3000
S-dof	c	0.2049	0.4496	0.4519	0.6650	0.6865	69.5000
	e	0.2030	0.4471	0.4504	0.6789	0.7011	74.4000
	o	0.2012	0.4445	0.4434	0.7143	0.7257	83.3000
Motion blur	c	0.2040	0.4488	0.4516	0.6989	0.7174	73.9000
	e	0.2041	0.4488	0.4510	0.6964	0.7151	74.6000
	o	0.2012	0.4442	0.4415	0.7111	0.7233	84.6000

color-harmony (*Color-H*), rule of thirds (*Rule of thirds*), shallow depth of field (*S-dof*), and motion blur (*Motion blur*) features are presented in Table 6 and 7. For convenience, f represents the fusion type, and c , e , and o are abbreviations for *Concat*, *Eltwise* and *Ours*, respectively.

In addition to experiments on AADB and AVA datasets that already contain the five aesthetic labels, we also performed experiments on the CUHK dataset without these aesthetic labels. CUHK dataset contains 29,690 images in total. The images were obtained from dpchallenge website. CUHK is divided into two categories: high quality images and low quality images. High-quality images are marked as 1 and low-quality images are marked as 0. In fact, 10,000 images are randomly selected as our experimental data. These data are called mini CUHK dataset (mCUHK). The number of training, validating, and testing sets are 8,500, 500

TABLE 9. Compared with other methods on the AADB dataset.

Methods (Years)	$Sp_{-}\rho$ ↑	Acc(%) ↑
AADB [31] (2016)	0.6782	-
Ours	0.6846	76.6000

TABLE 10. Compared with other methods on the AVA dataset.

Methods (Years)	Time(s)	$Sp_{-}\rho$ ↑	Acc(%) ↑
S. Dhar et al. [27] (2011)	-	-	64.35
AVA [30](2012)	-	-	68.00
W. Lou et al. [28] (2013)	0.03	-	68.79
DCNN [16] (2014)	2.40	-	73.25
RDCNN [16] (2014)	3.00	-	74.46
SPP [33] (2015)	0.60	-	72.85
DMA [29] (2015)	-	-	74.46
DMA_AlexNet_FT [29] (2015)	-	-	75.41
AADB [31](2016)	-	0.5581	77.33
STCNN [6] (2017)	-	-	75.13
A-lamp [32] (2017)	-	-	82.50
K. Schwarz et al. [34] (2018)	-	-	75.83
M. Kucer et al. [37] (2018)	-	-	81.95
ILGNet [36] (2019)	-	-	82.66
X. Zhang et al. [35] (2019)	-	-	81.81
Ours	0.02	0.6619	82.8250

and 1,000, respectively. Firstly, we extract the five aesthetic features from the CUHK by using the method described in Section III – B. Secondly, three fusion methods are used to fuse each aesthetic feature. The fusion results are listed in Table 8. As can be seen from the table, even on data that do not contain these aesthetic labels, the fusion of aesthetic features, according to our method, can still achieve good results.

E. COMPARE WITH OTHER METHODS

Table 9 and 10 show the results of the proposed network and method on the AADB dataset and AVA dataset for image aesthetics assessment. The classification accuracy of our method reached 76.60% and 82.8250% on the two datasets, respectively. In addition to comparing accuracy with other methods, we also give the average running time. In the test phase, our method only takes 0.02s to test a 224×224 image. It is obvious that our method achieves state-of-the-art results compared to other recently proposed methods.

V. CONCLUSION

In this paper, we design a two-stream network to calculate the aesthetic quality of images. The upper stream of the network is an improvement to the SEResNet-50, and the improved network can increase the performance of the model and extract deep CNN features. Five traditional algorithms for extracting aesthetic features are put forward. These algorithms can extract the brightness, color-harmony, rule of thirds, shallow depth of field and motion blur feature maps well. We propose the feature fusion layer, which can fuse deep CNN features and aesthetic features without adding feature dimensions. The results show that the aesthetic score produced by the proposed method is similar to the result of manual scoring. In addition,

this technique achieves higher classification accuracy (Acc) than existing methods.

REFERENCES

- [1] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1531–1544, Jul. 2019.
- [2] L. Baraldi, C. Grana, and R. Cucchiara, "Scene-driven retrieval in edited videos using aesthetic and semantic deep features," in *Proc. ACM Int. Conf. Multimedia Retr. (ICMR)*, 2016, pp. 23–29.
- [3] I. Arora and N. K. Garg, "Bi-featured image quality assessment with the hierarchical image quality enhancement algorithm," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Aug. 2016, pp. 1–6.
- [4] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.
- [5] Y. Zhang and D. M. Chandler, "Opinion-unaware blind quality assessment of multiply and singly distorted images via distortion parameter estimation," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5433–5448, Nov. 2018.
- [6] Y. Kao, R. He, and K. Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1482–1495, Mar. 2017.
- [7] H.-J. Lee, K.-S. Hong, H. Kang, and S. Lee, "Photo aesthetics analysis via DCNN feature encoding," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1921–1932, Aug. 2017.
- [8] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 80–106, Jul. 2017, doi: [10.1109/MSP.2017.2696576](https://doi.org/10.1109/MSP.2017.2696576).
- [9] L. Marchesotti, F. Perronnin, and F. Meylan, "Learning beautiful (and ugly) attributes," in *Proc. BMVC*, 2013, pp. 1–11. [Online]. Available: <http://www.bmva.org/bmvc/2013/Papers/paper0007/paper0007.pdf>
- [10] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1784–1791.
- [11] S.-M. Huang and J.-F. Yang, "Linear discriminant regression classification for face recognition," *IEEE Signal Process. Lett.*, vol. 20, no. 1, pp. 91–94, Jan. 2013.
- [12] S. Saleem and R. Sablatnig, "A robust SIFT descriptor for multispectral images," *IEEE Signal Process. Lett.*, vol. 21, no. 4, pp. 400–403, Apr. 2014.
- [13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 25. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105.
- [14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9, doi: [10.1109/CVPR.2015.7298594](https://doi.org/10.1109/CVPR.2015.7298594).
- [15] Z. Dong, X. shen, H. Li, and X. Tian, "Photo quality assessment with DCNN that understands image well," in *Proc. Int. Conf. Multimedia Model.*, 2015, pp. 524–535.
- [16] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "RAPID: Rating pictorial aesthetics using deep learning," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 457–466.
- [17] D. Harman, *The Digital Photography Handbook*, 4th ed. London, U.K.: Quercus, 2012.
- [18] B. Barnbaum, *The Art of Photography: The Essence of Photography Seeing and Creativity*. Beijing, China: The People's Posts and Telecommunications Press, 2017.
- [19] Y. Tan, P. Tang, Y. Zhou, W. Luo, Y. Kang, and G. Li, "Photograph aesthetical evaluation and classification with deep convolutional neural networks," *Neurocomputing*, vol. 228, pp. 165–175, Mar. 2017.
- [20] M. Kairanbay, J. See, L.-K. Wong, and Y.-L. Hii, "Filling the gaps: Reducing the complexity of networks for multi-attribute image aesthetic prediction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 3051–3055.
- [21] X. Fu, J. Yan, and C. Fan, "Image aesthetics assessment using composite features from off-the-shelf deep models," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3528–3532.
- [22] X. Tian, Z. Dong, K. Yang, and T. Mei, "Query-dependent aesthetic model with deep learning for photo quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2035–2048, Nov. 2015.
- [23] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2008, pp. 304–306.
- [24] R. Jain, R. Kasturi, and B. G. Schuck, *Machine Vision*. New York, NY, USA: McGraw-Hill, 1995.
- [25] X. Wang, "Laplacian operator-based edge detectors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 886–890, May 2007.
- [26] L. G. Roberts, *Machine Perception of Three-Dimensional Solids*. Cambridge, MA, USA: MIT Press, 1963, pp. 99–157.
- [27] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. CVPR*, Providence, RI, USA, Jun. 2011, pp. 1657–1664, doi: [10.1109/CVPR.2011.5995467](https://doi.org/10.1109/CVPR.2011.5995467).
- [28] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1930–1943, Dec. 2013, doi: [10.1109/TMM.2013.2269899](https://doi.org/10.1109/TMM.2013.2269899).
- [29] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 990–998.
- [30] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2408–2415.
- [31] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 662–679.
- [32] S. Ma, J. Liu, and C. W. Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 722–731.
- [33] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
- [34] K. Schwarz, P. Wieschollek, and H. P. A. Lensch, "Will people like your image? Learning the aesthetic space," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Tahoe, NV, USA, Mar. 2018, pp. 2048–2057.
- [35] X. Zhang, X. Gao, W. Lu, and L. He, "A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2815–2826, Nov. 2019.
- [36] X. Jin, L. Wu, X. Li, X. Zhang, J. Chi, S. Peng, S. Ge, G. Zhao, and S. Li, "ILGNet: Inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation," *IET Comput. Vis.*, vol. 13, no. 2, pp. 206–212, Mar. 2019.
- [37] M. Kucer, A. C. Loui, and D. W. Messinger, "Leveraging expert feature knowledge for predicting image aesthetics," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5100–5112, Oct. 2018, doi: [10.1109/TIP.2018.2845100](https://doi.org/10.1109/TIP.2018.2845100).
- [38] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Computer Vision*. Berlin, Germany: Springer, 2006, pp. 288–301.
- [39] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 419–426.
- [40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141, doi: [10.1109/CVPR.2018.00745](https://doi.org/10.1109/CVPR.2018.00745).
- [42] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.-Q. Xu, "Color harmonization," *ACM Trans. Graph.*, vol. 25, no. 3, p. 624, Jul. 2006, doi: [10.1145/1141911.1141933](https://doi.org/10.1145/1141911.1141933).



XUEWEI LI received the bachelor's and master's degrees from the Xi'an University of Technology. She is currently pursuing the Ph.D. degree with the Beijing University of Posts and Telecommunications. Her current research interests include image aesthetic assessment, image processing, and machine learning.



XUEMING LI received the B.E. degree in electronics engineering from the University of Science and Technology of China, in 1992, and the Ph.D. degree in electronics engineering from the Beijing University of Posts and Telecommunications (BUPT), in 1997. From 1997 to 1999, he was a Postdoctoral Researcher with the Institute of Information Science, Beijing Jiaotong University. He has been with BUPT, since 1999. In 2002, he was a Guest Lecturer with Karlsruhe University,

Germany. To date, he has undertaken many state and enterprise research and development projects, and he has authored or coauthored three books and over 50 articles in the field of multimedia information processing and transmission. His current research interests include digital image processing, video coding, and multimedia telecommunication. He is a Senior Member of the Chinese Institute of Electrics and the China Society of Image and Graphics.



GANG ZHANG received the bachelor's and master's degrees from Xi'an Polytechnic University and Xidian University, respectively. He is currently pursuing the Ph.D. degree with Space Engineering University. His research interests include computer vision, image pattern recognition, and deep learning.



XIANLIN ZHANG received the master's degree from the Institute of Information Engineering, Northeastern University, in 2013, and the Ph.D. degree from the Beijing University of Posts and Telecommunications, in 2019. She is currently a Postdoctoral Researcher with the School of Computer, Beijing University of Posts and Telecommunications. Her research interests include sketch-based image retrieval, image pattern recognition, and deep learning.

...