Virtual Reality & Intelligent Hardware

• Article •

February 2023 Vol. 5 No. 1: 42–56
10.1016/j.vrih.2022.06.006

IAACS: Image aesthetic assessment through color composition and space formation

Bailin YANG[1*], Changrui ZHU[1], Frederick W. B. LI[2], Tianxiang WEI[3], Xiaohui LIANG[4], Qingxu WANG[1]

1. *School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou* 310018*, China;*
2. *Department of Computer Science, University of Durham, DH1 3LE, United Kingdom;*
3. *College of Intelligence and Computing, Tianjin University, Tianjin* 300072*, China;*
4. *State Key Laboratory of Virtual Reality Technology and Systems, Beihang University , Beijing* 100191*, China*

Received 9 March 2022; **Revised** 5 June 2022; **Accepted** 23 June 2022

Abstract: Background Determining how an image is visually appealing is a complicated and subjective task. This motivates the use of a machine-learning model to evaluate image aesthetics automatically by matching the aesthetics of the general public. Although deep learning methods have successfully learned good visual features from images, correctly assessing the aesthetic quality of an image remains a challenge for deep learning. **Methods** To address this, we propose a novel multiview convolutional neural network to assess image aesthetics assessment through color composition and space formation (IAACS). Specifically, from different views of an image—including its key color components and their contributions, the image space formation, and the image itself—our network extracts the corresponding features through our proposed feature extraction module (FET) and the ImageNet weight-based classification model. **Result** By fusing the extracted features, our network produces an accurate prediction score distribution for image aesthetics. The experimental results show that we have achieved superior performance.

Keywords: Image aesthetic assessment; Color composition; Space formation; Deep learning

Supported by the National Key R&D Program of China (No: 2018YFB1403202) and the National Natural Science Foundation of China (62172366).

Citation: Bailin YANG, Changrui ZHU, Frederick W. B. LI, Tianxiang WEI, Xiaohui LIANG, Qingxu WANG. IAACS: Image aesthetic assessment through color composition and space formation. Virtual Reality & Intelligent Hardware, 2023, 5(1): 42–56

1 Introduction

The automatic assessment of image aesthetics is challenging. This relates to the development of image assessment methods that resemble the aesthetics of the public. Specifically, there are many factors for judging the aesthetic quality of an image, such as the colors constituting the image, how well those colors go together with each other, the content objects that constitute the image, and how they are organized in the image. The ability to extract and reason these factors is the key to assessing the image aesthetic quality. With the

*Corresponding author, ybl@zjgsu.edu.cn

development of deep learning technologies, researchers have introduced deep convolutional neural networks (CNNs) to perform image aesthetic evaluation[1]. Such methods can assist people in judging image quality, particularly those without rich image aesthetic knowledge and photography experience.

In recent years, deep CNNs have shown improvements in evaluating image quality and have become the mainstream method for solving image aesthetic assessment problems. The RAPID model[2] can be considered the first attempt to train CNNs using aesthetic data. They used an architecture similar to that of AlexNet, with the last fully connected layer producing 2-dimensional probabilities for aesthetic binary classification. Later, DMA-net[3] developed a deep multipatch aggregation network training method to allow multiple patches generated from an image to train the model. Kong et al. proposed learning aesthetic features by unifying photo-style attributes and content information to rate image aesthetics[4].Wang et al. proposed the brain-inspired deep network (BDN) by modifying the ResNet. It simulates the potentially complex neural mechanism of human perception, performing better than most standard aesthetic models with manual features and linear classifiers[5]. Ma et al. proposed an A-Lamp CNNs architecture to simultaneously learn fine-grained layouts and overall layouts[6]. Talebi et al. proposed using the squad EMD loss and predicted the distribution of scores as a histogram[7]. Li et al. proposed the use of manually extracted aesthetic features and original features to form an edge plus center network layer fusion method for aesthetic scoring[8]. Kang et al. proposed an explainable visual aesthetics (EVA) dataset that is expected to contribute to future research on understanding and predicting visual quality aesthetics[9]. Nascumento et al. believed that the perceived naturalness and preference were perceptually closely related and might be driven by related mechanisms. In other words, the more natural an image is, the more people like it[10].

Typically, performing a professional judgment of image aesthetic quality is challenging for ordinary users. A promising direction to address this gap is to develop machine-learning models for performing such tasks. This may support aesthetic image sorting, leading to better results in image retrieval, album organization, and photo editing[4,7]. In addition, Ma et al. greatly improved the attractiveness of multimedia search re- sults based on image aesthetic sorting[11]. Talebi et al. applied image aesthetic evaluation to tune a denoising filter to achieve better image quality[7].

In the field of art, color and composition are the two fundamental factors that determine the image quality. Lu et al. designed a conditional random field (CRF)-based color harmony model using a deep neural network to obtain coherence properties between image patch pairs, and embedded these relations along with each patch's own color harmony into a CRF[12]. Obrador et al. proposed low-level image composition features that approximated traditional photography composition guidelines and used these features to build an image aesthetics classifier[13]. Mai et al. presented a composition-preserving deep ConvNet method that directly learns aesthetic features from input images without any image transformations[14]. Zhao et al. proposed using MobileNetV2 to extract edge and global features by tuning these features through a gate unit[15].

Compared to previous models, there is still a lot of room for exploring aesthetic feature extraction. Inspired by the three basic disciplines of the Bauhaus educational model[16]—namely, plane composition, color composition, and three-dimensional composition, which are also the basic disciplines for many visual arts studies—we reason image aesthetics as color composition and space formation, propose extracting the color palette as the color position features, and extract image contour maps based on edge detection to formulate space formation features. We extracted these features using the proposed feature extraction module (FET). We also extracted image content features from an input image using the ImageNet-based[17] classification network model. By fusing the extracted features, we generated a ten-category score distribution to represent the image aesthetics. The main contributions of this study are as follows.

• We are the first to use color composition, space formation, and visual features of image contents to assess image aesthetics through deep learning.

• We propose the extraction of color composition features based on the color palette and the contribution of each color component. We also propose the extraction of space formation features by analyzing the contours of the image content.

• We develop a novel multi-view learning model to extract and merge features and effectively evaluate image aesthetics.

2 Related work

2.1 Color palette

Color is fundamental in image aesthetics. A key research question is how colors can be extracted from an image to reflect the color composition effectively. One solution is to make use of a color palette, which is an unordered list of colors, as illustrated in Figure 1. Color palettes can be used to define the color space of a display device or to represent the main colors used in an image. Lo et al. extracted the dominant colors of an image as a color palette to assess the image aesthetics[18]. Alternatively, color palettes have also been used as a factor to assess the prices of paintings[19].

Figure 1 The traditional five-color palette typically does not consider how each color contributes to an image.

In online communities, such as https://www.colourlovers.com and https://www.color.adobe.com, five-color palettes are widely used to collect and rank user preferences for color combinations. They serve as valuable sources of color recommendations for designers and the general public. Typically, a color palette comprises a list of colors of different types, in which each color contributes equally to the palette. In addition to reflecting the major colors constituting an image, the color palette is a tool for assessing color harmony. Karp et al. designed 12 basic color wheels, which are regarded as the most basic color harmony templates[20]. These color wheel harmony template groups play a significant guiding role in color design[21]. O'Donovan et al. proposed a scoring study using color palettes and created a data-driven LASSO regression model to evaluate the degree of harmony of color combinations based on five-color palettes[22]. Tan et al. performed image decomposition based on the RGBXY space to extract color palettes from images and investigate color harmony[23]. Lu et al. designed a CRF based color harmony model[12]. Schloss et al. aimed to characterize user aesthetic responses to color pairs, considering the colors that users prefer in combination and how the spatial organization of component colors influences pair preference[24]. Therefore, they provide possible phenomenological and ecological explanations for the reported results. This reflects the importance of color in image aesthetics. Compared with the color wheel, the color palette can summarize the color combination contained in an image more intuitively. We are the first to propose using a color palette and the contributions of its color components for assessing image aesthetic.

Some common methods for generating color palettes include: 1) applying a clustering algorithm to classify the corresponding color values of image pixels, 2) learning from the color palette extracted by professional designers and extracting them manually according to rules, and 3) using ready-made palettes from the color palette dataset website.

2.2 Space formation

Originating from the design curriculum reform of the Bauhaus College in Germany in 1919[16], color, space, and composition are defined as three basic components that provide rational and logical reasoning of art and design quality. Specifically, space formation is based on the three-dimensional arrangement of visual elements

that constitute a visually appealing space. Sneum et al. developed a theoretical system of composition art and laid a good foundation for the concept of space formation, stating that a pleasant space is usually realized by specific arrangements of points, lines, and planes[25]. That is, the variations in the shapes involved and their arrangements induce different visual feelings. Therefore, the ability to extract and quantify the space formation characteristics of an image is important for aesthetic assessment.

Obrador et al. proposed low-level image composition features that approximated the traditional photography composition guidelines[13]. Mai et al. applied adaptive spatial pooling to facilitate image composition feature preservation, avoiding the input image distortion induced by fixed-size processing requirements[14]. Zhao et al. proposed using a fully connected graph to characterize image composition. These methods are complicated in architecture and do not achieve satisfactory results[15]. By contrast, we propose using image edge detection to extract important structural image attributes that reflect the space formation of an entire image. The Canny operator or its variants are the most commonly used edge detection methods[26], which design an optimal pre-smoothing filter for edge detection and are optimized by a first-order Gaussian derivative kernel. The Sobel edge detection algorithm[27] is simple, efficient, and widely used in practical applications, although it is less accurate than the Canny operator.

2.3 Traditional image aesthetic assessment methods

These methods artificially design feature extractors, which require many engineering technologies and domain expertise, and evaluate image aesthetics using machine learning methods. Tong et al. experimentally identified a set of global low-level features, including blur, contrast, sharpness, and saliency, to model the global aesthetic aspects of images and classify photographs taken by professional photographers and ordinary users[28]. Datta et al. combined low-level and high-level features to train support vector machine (SVM) classifiers for the binary classification of image aesthetic quality[29]. Ke et al. proposed the use of global edge distribution, color distribution, tone count, and contrast and brightness indicators as image features, and trained a naive Bayes classifier based on these features[30]. Liu et al. proposed a semi-supervised deep active learning algorithm for aesthetic evaluation that reduced the inherent label noise of pictures[31].

Image aesthetic assessment is broadly categorized into three different approaches. The first one is the two-category prediction of image aesthetics, which produces a two-category judgment of an image through some image aesthetics assessment algorithms, judging whether the image is visually appealing or not. The second one is score prediction. This approach applies certain image aesthetic assessment algorithms to produce a score to judge image aesthetic, typically ranging from 1 to 10 points. The final one is scoring distribution prediction. Similar to the second approach, this approach still ranks the image aesthetic from 1 to 10 points, yet producing a scoring distribution instead of a fixed value for rating an image. These three types of image aesthetic assessment methods gradually developed and progressed in order. When there is a prediction of the rating distribution of an image, it can easily calculate the specific rating value and the binary classification for this image, i.e., whether it is good or bad. Similarly, given the rating value of an image, it can easily obtain the binary classification of that image.

2.4 Deep learning methods of aesthetic image

Learning image features from a large amount of data has shown high performance in recognition, positioning, retrieval, and tracking tasks, surpassing methods based on handcrafted features. Since Krizhevsky et al. applied CNNs for image classification, many researchers have started to learn image representation via deep learning methods[32]. The RAPID model proposed by Lu et al. was the first attempt to train CNNs using aesthetic data[2]. Wang's BDN simulated the perception network of the human brain and achieved satisfactory results[5]. A-Lamp proposed by Ma et al. considered the granularity and global characteristics[6]. These

methods only focused on the simple binary classification of image aesthetics. In contrast, the proposed method can rank image aesthetics on multiple levels. An initial work was proposed in NIMA by Talebi et al.[7], which replaced the last layer of a CNN classifier network with a fully connected layer of 10 neurons, followed by soft-max activation to produce a distribution of image aesthetic rankings. No specific image composition features were trained in the model. By using DenseNet-121 as the backbone, Liu et al. proposed the integration of low- and high-level features to preserve visual details for extracting aesthetic properties properly and applied atrous spatial pyramid pooling (ASPP) to encode multi-scale features that capture a diverse range of image contexts[33]. Zhao et al. proposed combining both composition and aesthetic features and used a gate unit to tune these features[15]. Li et al. proposed the use of manually extracted aesthetic features and original features to perform an edge plus center network layer fusion method for aesthetic scoring[8]. None of these three methods considered the two most important factors affecting image aesthetics, namely, color composition and space formation. Lyu et al. proposed a novel user-guided personalized image aesthetic assessment framework that leveraged user interactions to retouch and rank images for aesthetic assessment using deep reinforcement learning (DRL)[34]. It generated a personalized aesthetic distribution that was more in line with the aesthetic preferences of different users. Chambe et al. evaluated whether certain computational models of aesthetics were generalized and performed well over professional photographs. The models were then fine-tuned using professional photographs[35]. Sheng et al. presented a multi-patch (MP) aggregation method for image aesthetic assessment. It also proposed a set of objectives with three typical attention mechanisms, namely, average, minimum, and adaptive, and evaluated their effectiveness on the aesthetic visual analysis (AVA) benchmark[36]. Sheng et al. also revisited the problem of image aesthetic assessment from a self-supervised feature learning perspective and designed two novel pretext tasks to identify the types and parameters of editing operations applied to synthetic instances[37]. The features from their pretext tasks were then adapted for a one-layer linear classifier to evaluate performance in terms of binary aesthetic classification.

3 Our proposed IAACS model

Inspired by the design curriculum reform of the Bauhaus College[16], we propose image aesthetic evaluation as a combined assessment of color composition, space formation, and image visual content. Considering both the contribution of each of these features and their combined contribution to image aesthetics, we developed a multi-view learning model to learn these features for rating image aesthetics, as shown in Figure 2.

For color composition, we apply the k-means clustering algorithm[38] to analyze the color features of an image and obtain five key color patches, together with their contributions. For space formation, we exploit

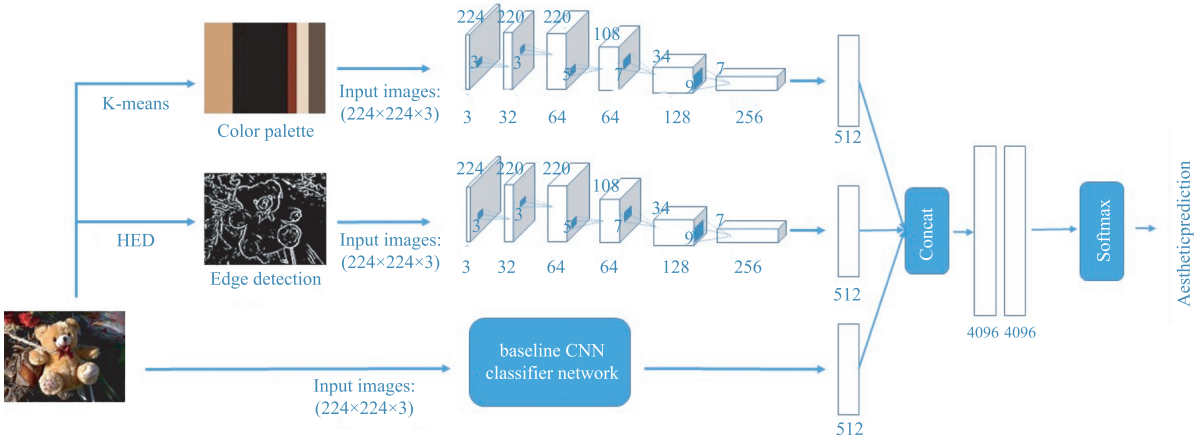

Figure 2 Overall architecture of our proposed IAACS model.

edge detection to extract the contours of the image content for processing. Specifically, we applied the holistically nested edge detection (HED) algorithm[39] to extract multi-scale and multi-level contour features from an image for our model to train the space formation. HED performs image-to-image prediction using a deep learning model that leverages fully convolutional neural networks and deeply supervised nets. HED tackles two important issues in vision problems: 1) holistic image training and prediction, and 2) multi-scale and multi-level feature learning. To train each of the color composition and space formation features, we separately processed them using our proposed FET. Additionally, we learn the image visual features perceived by humans through a baseline CNN classifier network as another task. Finally, the features extracted from the above tasks were fused through two fully connected layers. We then applied softmax to obtain the score distribution of the aesthetic image.

3.1 Color composition feature

Our method uses the AVA dataset[40], which provides aesthetic annotations of images as input to train our model for image aesthetic scoring. Aesthetic annotations provide only a score for each image to indicate the visual quality. Such a score not only reflects how well the image contents are perceived by users, but also encapsulates the quality of the types of colors being used. However, this color quality may not be directly correlated with the annotated image scores. This motivates us to learn image aesthetic scoring by considering the color composition and the contribution of each color component in an image.

Color palettes are a popular choice for representing the main color characteristics of an image. It comprises an unordered list of colors. By controlling the types and numbers of colors involved in a color palette, we can characterize image coloring in a coarse-grained or fine-grained manner. Similar to Yang et al.[41], our method constructed a five-color palette for each image. This is because most publicly available color palette dataset websites support five-color palettes. We then chose five-color palettes to form our color composition features. This widens the choice of compatible annotated image rating data for training our model. To construct an image color palette, we applied the k-means algorithm[38] to cluster the color pixels of an image. For the color space representation, we decided to use the LAB color model. While the LAB model can express some color range that the RGB color model cannot express, it is also a better model, as it is formulated based on physiological characteristics and is device-independent. Therefore, we converted all input images into the LAB color space before further processing. To extract the major color composition of an image, we used the LAB color value of each image pixel as the initial sample and randomly selected five initial cluster centers based on them. We then repeatedly applied clustering operations by setting k=5 until the optimal result was reached.

To strengthen the image color representation, we considered the contribution of each color component of the palette of an image. Hence, after obtaining the optimal clustering results of the image color space, we calculated the image area ratio contributed by each cluster center. As shown in Figure 3, we eventually obtained a contribution-aware palette of five different color patches, with the area of each color patch being proportional to its contribution in the corresponding image.

After obtaining the contribution-aware color palette for each input image, we extracted the image color composition features. Owing to the excellent image processing capability of CNNs, we used a CNN to extract color composition features. This is because the color palettes contain only simple color and area information. For the effective extraction of these simple features, we attempted to use a network with a small number of layers because shallow networks can learn low-level features better. Therefore, we designed a FET consisting of five convolutional layers and one fully connected layer, as shown in Figure 4. The sizes of the convolution kernels of the five convolutional layers are 3×3, 3×3, 5×5, 7×7, and 9×9. The number of convolution kernels was 32, 64, 64, 128, and 256. The last layer, that is, the fully connected layer, was used to adjust the feature

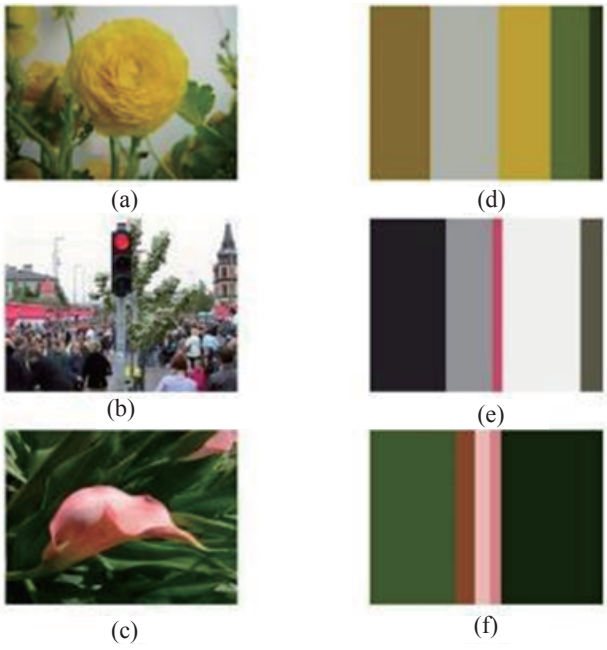

Figure 3 Extracted color palettes from sample images: (a–c) are the input images, with their corresponding extracted color palettes shown in (d–f).

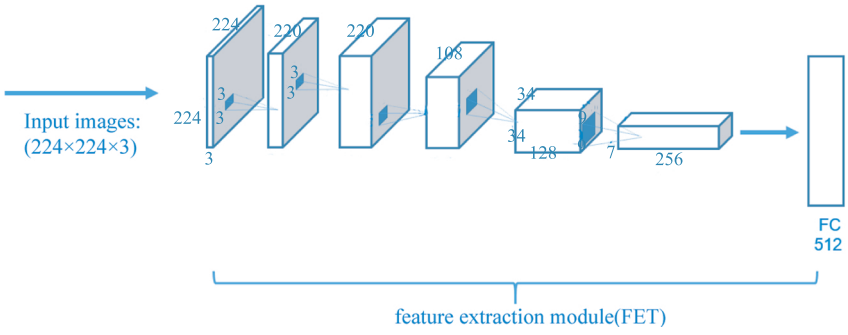

Figure 4 Feature extraction module (FET). This network architecture consists of five convolutional layers and one fully connected layer. It can well extract the type and size information of colors in a palette, as well as the holistic contour map obtained by holistically nested edge detection (HED)[39].

size for preparing the feature fusion operation that happened next.

3.2 Space formation feature

Space formation is another important factor that influences image aesthetics. As shown in Figure 5, edge detection allows us to effectively obtain the space formation features of an image. In addition, edge-detection maps reveal different levels of image composition. Therefore, we modified the HED method[39] to process input images and obtain holistic contour maps to extract proper structural attributes for learning image space formation.

Our aim is to obtain both the overall layout of the objects inside an image and the detailed contours of each object. As depicted in Figure 5, we examined the outputs from different side-output layers of the HED. Clearly, as shown in Figure 5(d–g), outputs from side-output layers 3 to 5 and the final fusion layer (side-output-fuse) generally preserve the overall and high-level object layouts. However, they all lack object details and can extract perceptually irrelevant layout details from a blurry background. These extracted attributes may not accurately reflect how people perceive objects and their layouts in the image. Hence, they are not suitable

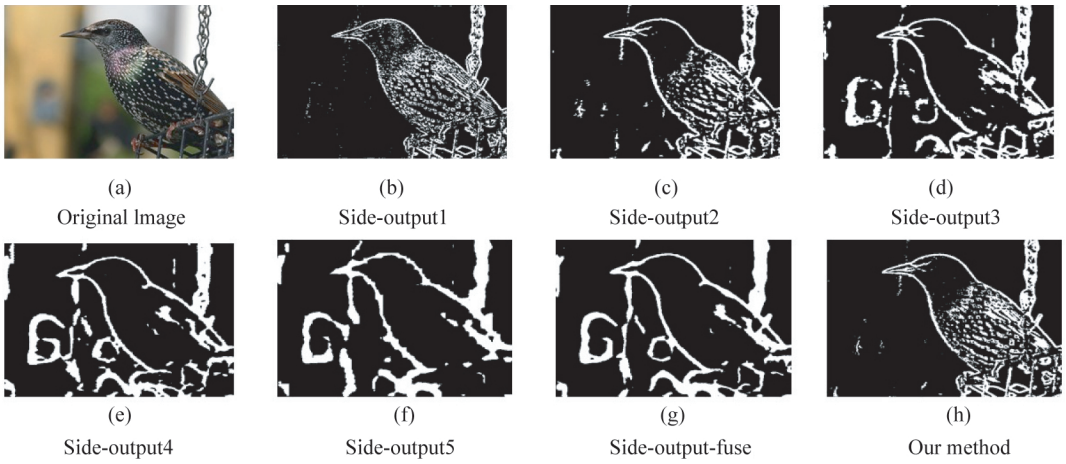

Figure 5 Example outputs from each network layer of HED and the holistic contour map generated in our method. (a) Example input image, (b)–(g) outputs from side-output layers 3 to 5 and the final fusion layer (side-output-fuse), respectively. (h) Our holistic contour map.

for learning about space formation. In contrast, it can be observed that side-output layer 1 (Figure 5b) provides very fine details of image objects while side-output layer 2 (Figure 5c) emphasizes more on the overall object layout. As shown in Figure 5h, we fused the outputs from these two layers to generate a holistic contour map by stacking, which can simultaneously preserve the overall object layout and avoid conveying excessive object details, forming a suitable input for learning the space composition.

After obtaining the holistic contour map, similar to learning color composition, we applied our proposed FET to learn space formation. Because a holistic contour map comprises only low-level contour features without complicated textures, a very deep network is not required for learning.

3.3 Feature extraction of input image content

In addition to the color and structural features obtained by color composition and space formation, we also learn the visual characteristics perceived by humans from an input image. Because visual features from images are significantly more complicated than those from color composition and space formation, a deeper network is required for feature learning. Typically, CNNs are suitable network architectures for this purpose. We compared the performance of typical CNN architectures on ImageNet and selected the three with better performance as our candidate basic backbone networks, namely, VGG16[42], InceptionResNetV2[43], and NASNet[44].

VGG16[42] comprises 13 convolutional layers, 3 fully connected layers, and 5 pooling layers. Its convolution filter size is small, that is, only 3×3, thus constructing a deep network. InceptionResNetV2[43] is based on an inception module[45]. It borrows the idea from ResNet[46] using residual connections[46] to add shortcuts to the model to allow training images from a much deeper network. NASNet[44] draws on the repeated stacking ideas of current mainstream excellent network structures, such as ResNet[46] and Inception[45], building the entire network structure by stacking convolution cells. The designed convolution cell mainly includes two types: 1) normal cell: convolution without changing the size of the input feature map, and 2) reduction cell: convolution that reduces the length and width of the input feature map to half of the original and reduces the size by increasing the size of the stride. The initial network weights used in our study were obtained by training the ImageNet dataset[17]. We discarded the remaining network layers in the baseline network and added a fully connected layer to prepare for the feature fusion. We used these network structures to extract the features from the input images.

3.4 Feature fusion and training

Finally, we integrate the learned features from the color composition, space formation, and visual features of the image content. Through the concept of feature optimization, for all the features we obtained, we divided them into small features by the number of channels and used the Concat method to perform feature fusion. In the last layer of each feature extraction we have added a fully connected layer to adjust the feature size. After the feature fusion layer, we added a global average pooling layer, two fully connected layers, each containing 4096 neurons, and a dropout layer. Subsequently, we added a 10-class softmax layer to predict image aesthetics, as shown in Figure 2.

In contrast to the traditional two-category image aesthetic prediction, we divided the image aesthetic score into 1 to 10 points, namely, from low to high aesthetic levels, producing a score distribution. The score distribution surpasses two-class image aesthetic prediction, precisely categorizing images into different levels of good (score >5) and bad (score≤5). The predicted aesthetic score of an image can be obtained by multiplying the score distribution by an array of 1 to 10. Moreover, the score distribution can also reflect whether there is a large difference in the evaluation of an image, where a standard deviation can be obtained for judgment.

We chose the earth mover's distance (EMD)[47] as the loss function for training. If we consider the distribution as two mounds with a certain amount of soil, then EMD is the minimum total work required to convert one mound into another. The workload is defined as the total amount of soil per unit multiplied by the distance it moves. It can be expressed as follows:

$$EMD(d,\hat{d}) = \left(\frac{1}{N}\sum_{K=1}^{N}\left|CDF_{d}(k) - CDF_{\hat{d}}(k)^{r}\right|\right)^{\frac{1}{r}} \tag{1}$$

where $d = \left[d(s_1), d(s_2), d(s_3), \cdots, d(s_N)\right]$ denotes the score distribution of ratings for the images.

Further, $\hat{d} = \left[\hat{d}(s_1), \hat{d}(s_2), \cdots, \hat{d}(s_N)\right]$, $s = s_1, s_2, \cdots, s_N$ denotes the ordered score, r is set to 2 to penalize the Euclidean distance between the CDFs, and $\mathrm{CDF}_d(k)$ is the cumulative distribution function.

4 Experiments

We conducted numerous experiments to prove the superiority of the proposed method. In this section, we introduce the dataset used. Second, we introduce our training details and conduct ablation studies to examine our method. Finally, we compared our method with previous methods to demonstrate the advantages of our proposal.

4.1 Dataset

The AVA dataset[40] is a database for aesthetic quality assessment, comprising approximately 255000 images collected at www.dpchallenge.com. Each image was voted by a number of people, with votes ranging from 78 to 549, and the average number of votes was 210. The voting scores ranged from 1 to 10. The higher the score, the better was the picture evaluation. AVA scores of >5 points were the majority. Moreover, the annotators included not only professional image workers and photographers, but also photography enthusiasts, which are more universal. Compared to other image aesthetic databases, AVA has a larger number of images and more detailed ratings. Therefore, we can directly use the voting scores from the AVA dataset as the score distribution for training, without additional processing. There are many unreal photographic images in the dataset as well as post-processed images, and many of the images are defective. We filtered the 255000 images, removed the defective images, and finally selected 249756 of them as our dataset.

4.2 Training details

For the 249756 images in the AVA dataset, we divided 50000 images into the validation and test sets, and all the remaining images were used for training. To extract the color palette, it was implemented in MATLAB. We implemented holistic contour map extraction and other network training processes using Keras in Python. The GPU used for training the network was an NVIDIA GV102. For the obtained color palettes and holistic contour maps as well as the original input images, we unified their sizes to 224×224×3, forming the inputs of our networks. For the network weights of the baseline CNN, we used the weights pretrained on the ImageNet dataset[17] to improve our training speed and training results. The batch size was set to eight. We also set the optimizer to Adam[48] with β_1=0.9 and β_2=0.99, and the initial learning rate was 1e-05. When the model's validation set loss did not decrease within 30 epochs, the learning rate was halved.

4.3 Ablation studies

To examine the effectiveness of different image features for image aesthetic assessment and how the use of different baseline networks affects the prediction accuracy, we designed three sets of ablation experiments. The first aimed to test the impact of adding color composition features and space formation features to image aesthetic evaluation. The second was designed to test the impact of different baseline network structures on the image aesthetic evaluation results. The last aimed to test how the correlation of low-level and high-level features between color composition and space formation affected the image aesthetic assessment.

Effectiveness of color composition and space formation features: NASNet[44] has absorbed the advantages of Inception[43] and ResNet[44] and offers the advantage of automatically selecting the network layer structure. These features cause NASNet to perform significantly better than the previous networks. In our work, we need to perform image assessment and feature extraction from the aesthetic aspect of images, where aesthetic features are often abstract and difficult to extract. Because NASNetlarge can learn to build a network layer structure through the program itself, it can extract the aesthetic features of images more fully. Because there is a network structure of NASNetlarge that has been pre-trained on ImageNet on keras, we used NASNetlarge as the baseline network structure to perform image aesthetic evaluation.

As presented in Table 1, we obtained results based on different settings, including training by color composition features, space formation features, both color composition features and space information features, and neither color composition nor space formation features. We then compared and analyzed these four results. From Table 1, we can see that among the four results, the network structure considering both color composition and space formation performed significantly better, which verifies the validity of our proposed idea. The color position and space formation of images can effectively guide image aesthetic assessment, and these two features are complementary and indispensable.

Effectiveness of baseline network structures: We used VGG16[42], InceptionResNetV2[43], and NASNetlarge[44] as the three baseline network structures to compare the image aesthetic evaluation results. We also verified the importance of using ImageNet's CNN pretraining model. The comparison results are presented in Table 2. We can see that ImageNet's CNN pre-training model has a great impact on the results. If it is not used, the performance drops significantly. Similar to our aforementioned findings, NASNetlarge achieved the best results because it has a better extraction ability for abstract image aesthetic features.

Table 1 Performance comparison with different aesthetic features on AVA dataset

Model	Accuracy(2 classes)	PLCC(mean)	SRCC(mean)	EMD
No Color composition and No Space formation (NASNetlarge)	80.09%	0.6545	0.6375	0.0693
Color composition (NASNetlarge)	79.89%	0.6608	0.6467	0.0687
Space formation (NASNetlarge)	80.61%	0.6636	0.6514	0.0679
IAACS (NASNetlarge)	86.29%	0.8003	0.8286	0.0573

Table 2 Performance comparison with different baseline networks on AVA dataset

Model	Accuracy(2 classes)	PLCC(mean)	SRCC(mean)	EMD
IAACS (No Pre-trained model)	74.40%	0.3412	0.3241	0.0820
IAACS (VGG16)	79.25%	0.6113	0.6006	0.0696
IAACS (InceptionResNetV2)	79 52%	0 6062	0 5963	0 0706
IAACS (NASNetlarge)	86.29%	0.8003	0.8286	0.0573

How the correlation of low- and high-level features between color composition and space formation affects image aesthetic assessment: As color composition and space formation are comparable features in terms of their complexity, we attempt to examine how their relationship affects the accuracy of image aesthetic assessment by performing early and late fusions in our multiview learning. Note that the image visual content feature was not involved in this study because it is deeper and incomparable to the other two features.

In early fusion, we shared the parameters of the two FETs during training, which means that feature fusion between color composition and space formation starts from their low-level features up to high-level features. In late fusion, we do not share parameters between the two FETs and only fuse their learned features, which means that only high-level features between the color composition and space formation are set to correlate. The comparison results are shown in Table 3. It can be observed that the late fusion approach produced significantly better results, which justifies our network design. These results suggest that image aesthetic assessment relies more on the correlation of high-level features between color composition and space formation. In contrast, introducing low-level feature correlation confuses the image aesthetic assessment.

Table 3 Performance comparison with different feature fusion methods on AVA dataset

Model	Accuracy(2 classes)	PLCC(mean)	SRCC(mean)	EMD
Early Fusion (NASNetlarge)	83.83%	0.7442	0.7561	0.0638
Late Fusion (NASNetlarge)	86.29%	0.8003	0.8286	0.0573

4.4 Comparison with previous methods

Here, we compare our method with previous methods to demonstrate that our proposal achieves a better performance. To compare the performance, we focus on three image aesthetic quality tasks currently included in the research community: 1) image aesthetic two-class prediction, 2) image aesthetic score prediction regression, and 3) image aesthetic score distribution prediction. For the two-class prediction of image aesthetics, high-quality images generally have a predicted score higher than 5, whereas low-quality images have a predicted score lower than or equal to 5, and accuracy is selected as the evaluation criterion. The definition of accuracy is as follows:

$$Accuracy = \frac{TP+TN}{P+N} \tag{2}$$

where TP and TN refer to the number of high-quality and low-quality images that were correctly predicted. P and N refer to the sum of the quantities of all high- and low-quality images, respectively. For the regression of image aesthetic prediction, we chose the Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SRCC) as evaluation criteria. The definition of a PLCC is as follows:

$$r(PLCC) = \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}} \tag{3}$$

The PLCC between two variables is defined as the quotient of the covariance and standard deviation between the two variables. X and Y represent the two variables. The definitions of SRCC and PLCC are similar:

$$r(SRCC) = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \overline{S})^2}} \quad (4)$$

The difference between them is that R and S refer to the level of the value and not the value itself.

SRCC and PLCC reflect the degree of linear correlation between the real image aesthetic score and the predicted value. The larger the value, the higher is the correlation. For the image aesthetic score distribution prediction, we used EMD to measure the gap between the true value and the predicted value. The smaller the value, the more accurate is the predicted score distribution.

The comparison results are listed in Table 4. In addition to the first few typical models, we found other output score distribution methods for comparison purposes. It can be observed that our method exhibits a better performance. Our accuracy is two percentage points higher than that of the previous best model, and our PLCC and SRCC are both higher than those of previously proposed methods. However, our EMD performance is not optimal. This may be because our model is a multi-view model and has too many parameters, making the learning effect on the more detailed rating distribution not particularly good. Examples of aesthetic assessment results are shown in Figure 6. The images were sampled from the AVA dataset. Each had multiple aesthetic assessment scores. Therefore, we took the average of these scores as the ground truth score for each image. We also report the aesthetic assessment scores obtained from the existing methods for comparison. As can be observed, our method generates scores closer to the ground truth ones, indicating that our method outperforms existing methods.

Table 4 Performance comparison with state-of-the-art methods on AVA dataset

Model	Accuracy(2 classes)	PLCC(mean)	SRCC(mean)	EMD
Lu et al.[2]	75.12%	–	–	–
Wang et al.[5]	76.80%	–	–	–
Ma et al.[6]	81 70%	–	–	–
Talebi et al. [7]	81.51%	0.6360	0.6120	0.0500
Zhang et al.[49]	81.81%	0.7042	0.6900	0.0450
Liu et al.[33]	83.59%	–	–	–
Zhao et al.[15]	82.35%	0.7600	0.7480	–
Li et al.[8]	82.08%	0.6524	0.6461	–
Sheng et al.[36]	83.03%	–	–	–
Pei et al.[34]	85.10%	–	0.6920	–
IAACS (NASNetlarge)	86.29%	0.8003	0.8286	0.0573

Our method: 5.9394
Ground truth: 6.0713
Talebi's method: 5.7391
Zhao'method: 5.8126
Li's method: 5.8617

Our method: 5.3448
Ground truth: 5.3023
Talebi's method: 4.9136
Zhao'method: 5.0103
Li's method: 5.1342

Our method: 3.5187
Ground truth: 3.3208
Talebi's method: 3.6198
Zhao'method: 3.7154
Li's method: 3.8237

Figure 6 Score comparison of our method and other methods. The score is calculated from the resulting rating distribution.

5 Conclusion

In this study, we propose a method for image aesthetic evaluation and prediction based on color composition,

space formation, and visual features of image content. We proved that adding these features offers better results, that is, provides better reasoning to the human sense of image aesthetics. We propose extracting color palettes with the contributions of their individual color components taken into account. We also propose extracting space formation holistic image contours to acquire the structural layout and object details. This information, as well as the original image, was fed into our novel multiview model to learn color composition, space formation, and visual image features for aesthetic evaluation. Our experimental results showed that we achieved superior performance.

Although our method achieved good results, we still need to consider several aspects in future research. First, we think that the evaluation of image aesthetics is very subjective, and we hope to make it as objective as possible. We will investigate the use of individual user preferences to improve the image aesthetic evaluation. Second, for an image, the source equipment that captures the image is a good point to consider, and relates to how an image is formulated. For instance, we can collect and predict whether the camera and relevant tools used comprise better functionalities that influence image aesthetics.

Declaration of competing interest

We declare that we have no conflict of interest

References

1 Deng Y, Loy C C, Tang X. Image aesthetic assessment: an experimental survey. IEEE Signal Processing Magazine, 2017, 34(4): 80–106
DOI: 10.1109/msp.2017.2696576

2 Lu X, Lin Z, Jin H, Yang J, Wang J Z. Rating image aesthetics using deep learning. IEEE Transactions on Multimedia, 2015, 17(11): 2021–2034
DOI: 10.1109/tmm.2015.2477040

3 Lu X, Lin Z, Shen X H, Mech R, Wang J Z. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). New York, ACM, 2015, 990–998
DOI: 10.1109/iccv.2015.119

4 Kong S, Shen X H, Lin Z, Mech R, Fowlkes C. Photo aesthetics ranking network with attributes and content adaptation. In: Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016, 662–679
DOI: 10.1007/978-3-319-46448-0_40

5 Wang Z Y, Chang S Y, Dolcos F, Beck D, Liu D, Huang T S. Brain-inspired deep networks for image aesthetics assessment. 2016: arXiv: 1601.0415. https://arxiv.org/abs/1601.04155

6 Ma S, Liu J, Chen C W. A-lamp: adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA, IEEE, 2017, 722–731
DOI: 10.1109/cvpr.2017.84

7 Talebi H, Milanfar P. NIMA: neural image assessment. IEEE Transactions on Image Processing, 2018, 27(8): 3998–4011
DOI: 10.1109/tip.2018.2831899

8 Li X, Li X, Zhang G, Zhang X. A novel feature fusion method for computing image aesthetic quality. IEEE Access, 2020, 863043–63054
DOI: 10.1109/access.2020.2983725

9 Kang C, Valenzise G, Dufaux F. EVA: an explainable visual aesthetics dataset. Joint Workshop on Aesthetic and Technical Quality Assessment of Multimedia and Media Analytics for Societal Trends. Seattle, WA, USA. New York, ACM, 2020, 5–13
DOI: 10.1145/3423268.3423590

10 Nascimento S M C, Marit Albers A, Gegenfurtner K R. Naturalness and aesthetics of colors – Preference for color compositions perceived as natural. Vision Research, 2021, 185, 98–110
DOI: 10.1016/j.visres.2021.03.010

11 Ma N, Volkov A, Livshits A, Pietrusinski P, Hu H D, Bolin M. An universal image attractiveness ranking framework. 2018

12 Lu P, Yu J B, Peng X J. Deep conditional color harmony model for image aesthetic assessment. In: 2018 24th International Conference on Pattern Recognition (ICPR). Beijing, China, IEEE, 2018, 2845–2850
DOI: 10.1109/icpr.2018.8546328

13 Obrador P, Schmidt-Hackenberg L, Oliver N. The role of image composition in image aesthetics. In: 2010 IEEE International Conference on Image Processing. Hong Kong, China, IEEE, 2010, 3185–3188
DOI: 10.1109/icip.2010.5654231

14 Mai L, Jin H L, Liu F. Composition-preserving deep photo aesthetics assessment. In: 2016 IEEE Conference on Computer Vision and

Pattern Recognition. Las Vegas, NV, USA, IEEE, 2016, 497–506
DOI: 10.1109/cvpr.2016.60
15 Zhao L, Shang M, Gao F, Li R, Huang F, Yu J. Representation learning of image composition for aesthetic prediction. Computer Vision and Image Understanding, 2020, 199: 103024
DOI: 10.1016/j.cviu.2020.103024
16 Droste M. Bauhaus 1919-1933. Dansk produktion: Book Service I/S, Copenhagen, Berlin, 1990
17 Suts kever A, Hinton G. Imagenet classification with deep convolutional neural networks advances in neural information processing systems. 1097: 1105. Alex Krizhevsky Ilya Sutskever and Geoffrey E Hinton, 2012
18 Lo K Y, Liu K H, Chen C S. Assessment of photo aesthetics with efficiency. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). Tsukuba, Japan, IEEE, 2012, 2186–2189
19 Stepanova E. The impact of color palettes on the prices of paintings. Empirical Economics, 2019, 56(2): 755–773
DOI: 10.1007/s00181-017-1413-4
20 Karp A, Itten J. The elements of color. Leonardo, 1972, 5(2): 180
DOI: 10.2307/1572567
21 Cohen-Or D, Sorkine O, Gal R, Leyvand T, Xu Y Q. Color harmonization. ACM Transactions on Graphics, 2006, 25(3): 624–630
DOI: 10.1145/1141911.1141933
22 O'Donovan P, Agarwala A, Hertzmann A. Color compatibility from large datasets. ACM Transactions on Graphics, 2011, 30(4): 1–12
DOI: 10.1145/2010324.1964958
23 Tan J C, Echevarria J, Gingold Y. Efficient palette-based decomposition and recoloring of images via RGBXY-space geometry. ACM Transactions on Graphics, 2018, 37(6): 1–10
DOI: 10.1145/3272127.3275054
24 Schloss K B, Palmer S E. Aesthetics of color combinations. In: SPIE Proceedings of Human Vision and Electronic Imaging XV. San Jose, California, SPIE, 2010
DOI: 10.1117/12.849111
25 Sneum G. Teaching design and form. 1965
26 Canny J. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, PAMI-8(6): 679–698
DOI: 10.1109/tpami.1986.4767851
27 Sobel I, Feldman G. A 3x3 isotropic gradient operator for image processing. Pattern Classification and Scene Analysis, 1973, 1, 271–272
28 Tong H H, Li M J, Zhang H J, He J R, Zhang C S. Classification of digital photos taken by photographers or home users. In: Advances in Multimedia Information Processing-PCM 2004. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, 198–205
DOI: 10.1007/978-3-540-30541-5_25
29 Datta R, Joshi D, Li J, Wang J Z. Studying aesthetics in photographic images using a computational approach. In: Computer Vision-ECCV 2006. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, 288–301
DOI: 10.1007/11744078_23
30 Ke Y, Tang X O, Jing F. The design of high-level features for photo quality assessment. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, NY, USA, IEEE, 2006, 419–426
DOI: 10.1109/cvpr.2006.303
31 Liu Z, Wang Z, Yao Y, Zhang L, Shao L. Deep active learning with contaminated tags for image aesthetics assessment. IEEE Transactions on Image Processing2018, 1
DOI: 10.1109/tip.2018.2828326
32 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84–90
DOI: 10.1145/3065386
33 Liu D, Puri R, Kamath N, Bhattacharya S. Composition-aware image aesthetics assessment. In: 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass, CO, USA, IEEE, 2020, 3558–3567
DOI: 10.1109/wacv45572.2020.9093412
34 Lyu P, Fan J Q, Nie X X, Dong W M, Jiang X H, Zhou B, Xu M L, Xu C S. User-guided personalized image aesthetic assessment based on deep reinforcement learning. 2021
35 Chambe M, Cozot R, Le Meur O. Behaviour of recent aesthetics assessment models with professional photography. 2019
36 Sheng K K, Dong W M, Ma C Y, Mei X, Huang F Y, Hu B G. Attention-based multi-patch aggregation for image aesthetic assessment. In: Proceedings of the 26th ACM international conference on Multimedia. Seoul Republic of Korea, New York, NY, USA, 2018
DOI: 10.1145/3240508.3240554
37 Sheng K K, Dong W M, Chai M L, Wang G H, Zhou P, Huang F Y, Hu B G, Ji R R, Ma C Y. Revisiting image aesthetic assessment via self-supervised feature learning. 2019
38 MacQueen J. Classification and analysis of multivariate observations. 5th Berkeley Symp. Math. Statist. Probability, 1967
39 Xie S, Tu Z. Holistically-nested edge detection. International Journal of Computer Vision, 2017, 125(1-3): 3–18

DOI: 10.1007/s11263-017-1004-z

40 Murray N, Marchesotti L, Perronnin F. AVA: a large-scale database for aesthetic visual analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA, IEEE, 2012, 2408–2415
DOI: 10.1109/cvpr.2012.6247954

41 Yang B, Wei T, Fang X, Deng Z, Li F W B, Ling Y, Wang X. A color-pair based approach for accurate color harmony estimation. Computer Graphics Forum, 2019, 38(7): 481–490
DOI: 10.1111/cgf.13854

42 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. CoRR, 2014
DOI: 10.48550/arXiv.1409.1556

43 Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-ResNet and the impact of residual connections on learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 4278–4284
DOI: 10.1609/aaai.v31i1.11231

44 Zoph B, Vasudevan V, Shlens J, Le Q V. Learning transferable architectures for scalable image recognition. 2017

45 Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. CoRR, 2015
DOI: 10.48550/arXiv.1502.03167

46 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. 2015

47 Levina E, Bickel P. The earth mover's distance is the mallows distance: some insights from statistics. Proceedings Eighth IEEE International, 2001
DOI: 10.1109/ICCV.2001.937632

48 Kingma D P, Ba J. Adam: a method for stochastic optimization. 2014

49 Zhang X, Gao X, Lu W, He L. A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction. IEEE Transactions on Multimedia, 2019, 21(11): 2815–2826
DOI: 10.1109/tmm.2019.2911428