Image Aesthetic Assessment Driven by Multimodal Features

Yun Liu[1], Zhipeng Wen*[1], Sifan Li[1], Daoxin Fan[1] and Guangtao Zhai[2]

*Corresponding author
[1] College of Information, Liaoning University, Shenyang, Liaoning, China
[2] Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China

Abstract. Our goal is to promote an effective image aesthetic assessment (IAA) model. In the current Internet era, it has become easier to obtain the text description of an image. With the dual-modal support of image and text, the image aesthetic assessment model will further reflect its superiority. To this end, we design a multimodal feature-driven guided image aesthetic assessment model (MFD). Firstly, multi-modal features are extracted through the feature extraction sub-network, including image-driven aesthetic features and content features, as well as text-driven semantic features. Each feature captures the implicit characteristics of different levels of human brain object analysis. Secondly, these multi-modal features are combined to form multi-modal combination features that contain multiple characteristics. Finally, the obtained multi-modal are combined for aesthetic assessment prediction. Experimental results on public image aesthetic assessment databases demonstrate the superiority of our model.

Keywords: Image Aesthetics Assessment, Multimodal Features, Feature Extraction Sub-network.

1 Introduction

With the widespread use of social media [1] and e-commerce platforms [2], image aesthetic prediction has become crucial for image retrieval systems [3], recommendation systems, and product marketing to enhance user experience and improve product sales [4]. With the development of machine learning technology, image aesthetic analysis has attracted increasing attention. Aesthetic perception is a relatively subjective concept, and different individuals vary in different preferences for the same image, making it challenging to establish an objective evaluation standard [5].

The focus of image aesthetic prediction research is to build a model to automatically predict the aesthetic quality of images. The task involves analyzing various visual features such as composition, color, content, and texture in images [6] [7]. By extracting relevant features from images and leveraging machine learning techniques, researchers aim to build accurate models for predicting aesthetic quality. Traditional approaches for image aesthetic prediction typically rely on hand-crafted features [8]. These methods often fail to capture complex visual patterns and aesthetic concepts effectively. By

using deep learning techniques, researchers have gradually replaced hand-crafted features with learned features, leading to more accurate predictions [9]. Convolutional neural networks (CNNs) in capturing hierarchical visual patterns in images [10] present great promise. AlexNet and regression loss are used in some architecture to predict a moderate pattern of mean scores for images [11]. Moreover, Visual Geometry Group Networks (VGGNets) are also fine-tuned to apply to comprehend how humans rate an image in aesthetic ways [12]. Recently, for the use of processing multiple scales of images, adaptive spatial pooling is applied [13], which shows us how a multi-net of pretrained VGGs extracts features at multiple scales. As the research in the area deepens, ResNet CNNs are applied in the process of predicting the quality of photos [14]. Additionally, attention mechanism [15] and reinforcement learning [16] have been employed to conduct image aesthetic prediction tasks. In summary, it can be observed that a series of applicable methods and tools have been used in aesthetic prediction of images, which makes an unprecedented leap in the area than the hand-crafted era [17].

Despite considerable progress made in image aesthetic prediction tasks, there are still several challenges. One of the fundamental challenges is the subjectivity associated with aesthetics [18]. Aesthetic preferences vary across individuals, cultures, and preferences, making it difficult to develop a universally applicable evaluation metric [19]. Furthermore, there is a need for more modality content that provides sufficient information for image aesthetic prediction models [20]. Most of the above works are built based on a single modal content, which can provide a limited performance improvement [21]. Multimodality contents like text, video, and audio can play important roles in downstream tasks. Some recent research works have proven that the multimodal based models yield better performance than single-modal based methods [22] [23].

Motivated by the above works, we build an image aesthetic assessment method driven by multi-modal feature perceptions in human brain, in which a specific feature extraction sub-network is designed to extract different categories of features from different modal content. Specifically, we design an Aesthetic Feature Extraction Sub-network (AFES) and a Content Feature Extraction Sub-network (CFES) to extract aesthetic features and content features of images respectively, which can well explain low-level visual perception. For high-level semantic understanding, we design a Semantic Feature Extraction Sub-network (SFES) to extract semantic features from text content. In order to fuse multiple types of features, a feature conversion (FV) module is built to change the dimensions and sizes of different features. Then the potential correlations of different categories of features are mined through Multi-Layer Perceptron (MLP), and finally the aesthetic assessment prediction distribution that correlates multiple types of features is obtained.

Our contributions mainly include the following aspects:

Based on human aesthetic perception characteristics, we design multi-modal feature extraction sub-network based on the image-driven aesthetic features and content features, and text-driven high-level semantic features to complementary the above visual features, which can boost the overall performance.

Considering the mutual influences between visual features and textual word, our model can effectively capture the relationship between different modes and retain the

characteristics of each mode by designing a feature conversion (FV) module, which can achieve accurate prediction.

The superiority performance of our model is demonstrated through experimental results on a public image aesthetic assessment database, which can prove the rationality of fusion visual features from image and textual word.

2 Related Works

Image aesthetic quality evaluation model can be categorized into two types: subjective evaluation and objective evaluation models. Subjective evaluation relies on human opinions to rate, which is costly and time consuming. Objective evaluation models attempt to propose a model to simulate the human perception to automatically assess the aesthetic quality of an image, which is of low cost and fast, and arouses more attention [24]. In early research, objective evaluation is proposed based on hand-crafted features to conduct image aesthetic quality classification. For example, Datta et al. [25] trained the SVM classifier based on some low-level features to classify the aesthetic quality, which are extracted according to photography rules. Wong et al. [26], based on the relationship between subject and background, and global features, built an aesthetic quality classification model. These features were extracted by detecting salient areas of the image. Overall, handcrafted features have clear physical meaning, and they tend to focus on specific physical features and struggle to fully capture the vital information of image aesthetics, which is limited its development.

With the rapid development of deep learning, the models based on CNN are increasingly built in image processing and have reliable results in predicting quality of image [27]-[30]. Compared with traditional handcrafted feature-based methods, these works have better performance in image perception assessment due to its ability to perform pixel-level quality assessment [28] [29]. The impressive results achieved by CNN provide us with a new way to evaluate the aesthetic quality assessment of images. By using patches randomly cropped from the original image, Lu et al. [31] proposed an architecture to categorize image aesthetic quality. Kao et al. [32] built a three independent convolutional neural networks model to extract objects, scenes, and textures for aesthetic learning. Mai et al. [33] proposed a model that did not require any image transformation, whose method is to use adaptive spatial convolution layers to train the model. In order to evaluate photos' quality, Zeng et al. [34] present a method to retrain AlexNet and ResNet CNNs. In order to learn the ranking correlation between two input images, Kong et al. [35] proposed an aesthetics ranking network based on Alex Net. By using the saliency characteristics of image, Ma et al. [36] proposed a layout aware framework to predict aesthetic scores. Considering the useful of multi-task learning, Li et al. [37] proposed a multi-task learning framework to predict the aesthetic evaluation. By using a universal aesthetic knowledge base, Li et al. [38] propose a new Knowledge Embedding model for image aesthetics assessment. She et al. [39] uses a novel graph convolution method to extract important image features to evaluate image aesthetic, which has improved the accuracy. Hosu et al. [40] and He et al. [41] simultaneously considered multi-level aesthetic features from a single modality and obtained good

results. Li et al. [42] fully considered the image visual attributes and proposed a TAVAR model, which yielded a high classification accuracy.

Although the above valuable studies have been previously conducted, image aesthetics assessment is an extraordinarily complex task. It is hard to continue to improve the overall performance by relying only on a single modality. Some recent research works have been built based on multimodal contents, which prove that multimodal-based methods can boost the result than single-modal based models **[22] [23]**. **For example,** Zhang et al. [43] proposed an image retrieval model based on visual features and text features. Taking audio modality into consideration, Wu et al. [44] combined audio and visual features proposed an action recognition method. To improve the classification results of images, He et al. [45] built a model that combines visual content and text information, which proves the nationality of visual and textual combination. Zhang et al. [46] encoded multimodal feature interactions by using MFB pooling method. To conduct multi-dimension aesthetic analysis, Miao et al. [47] propose a stacked multimodal co-transformer module. **Later,** Zhu et al. [48] proposed a new image-text interaction network (ITIN) to multimodal sentiment analysis. **Motivated by the above works,** Li et al. [1] proposed a multimodal Network that extracts more discriminative aesthetic representations by utilizing visual and text contents. The above works provide us a new way to conduct the image aesthetics assessment by utilizing multi-modal features.

3 Methodology

In this section, we introduce our proposed Multimodal Feature-Driven (MFD) image aesthetic assessment method. The specific structure of our model is present in Fig. 1. Firstly, the two different modal information, image, and text, are preprocessed separately. The image is resized to 224×224×3 and regularized using ImageNet's [10] to obtain a tensor. The text uses the pre-trained tokenizer [49] to convert the text into the corresponding token. Then, corresponding feature extraction sub-networks are designed for the preprocessed tensors and tokens to extract features of different modal information. Specifically, the Aesthetic Feature Extraction Sub-network (AFES) is designed for image information to extract aesthetic features and the Content Feature Extraction Sub-network (CFES) is designed to extract content features, and the Semantic Feature Extraction Sub-network (SFES) is designed for textual information to extract high-level semantic features. The feature conversion (FV) module within the feature extraction sub-network converts each extracted multi-modal feature into the same dimension. Finally, we splice three different types of features and use the Multi-Layer Perceptron (MLP) for aesthetic distribution prediction. Based on the aesthetic distribution prediction, the final aesthetic score $\hat{Y}$ can be calculated.

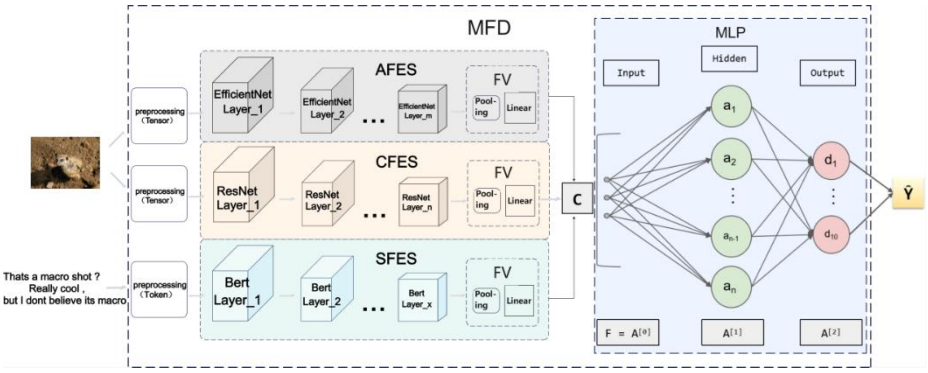

Fig. 1. Overall framework diagram of the proposed MFD

3.1 Feature Extraction Sub-network

Aesthetic Feature Extraction Sub-network (AFES). The basic expression of the aesthetic characteristics of images is obscure, and mining the potential connections of aesthetics in images is a necessary step in the image aesthetics assessment method. Therefore, aesthetic features, as the main features for judging the aesthetic quality of images, are also the core of the entire network. We design and use the efficient EfficientNet_v2_s [52] as the backbone of the network to mine the implicit expressions of image aesthetic features. Specifically, the fully connected layer classifier of the last layer of EfficientNet_v2_s is removed to establish an aesthetic feature extraction sub-network and initialize it using parameters pretrained on ImageNet. For the input image, the aesthetic feature extraction sub-network performs coding to extract a set of aesthetic feature tensors $AF^{c}_{b\times w\times h}$. The aesthetic feature tensors can be defined as follows:

$$AF^{c}_{b\times w\times h} = concat\{af^{1}_{b\times w\times h},\ af^{2}_{b\times w\times h},\ \ldots,\ af^{c}_{b\times w\times h}\} \tag{1}$$

Where, $af^{i}_{b\times w\times h} \in R^{b\times w\times h}$ represents an aesthetic feature tensor in the group of aesthetic feature tensors, and stores the data in the form of a three-dimensional matrix; $w \times h$ represents the width and height of the tensor, with a value of 7×7; c belongs to the number of channels, with a value of 1280; $AF^{c}_{b\times w\times h} \in R^{b\times c\times w\times h}$ represents the tensor of the group Store data in the form of a four-dimensional matrix.

In order to better combine with other features, we change the size and dimension of $CF^{c}_{b\times w\times h}$ through a feature conversion (FV) module. The process is defined as follows:

$$AF^{c}_{b\times 1\times 1} = meanpool(AF^{c}_{b\times w\times h}) \tag{2}$$

$$AF^{c}_{b} = squeeze(AF^{c}_{b\times 1\times 1}) \tag{3}$$

$$AF^{s} = linear(AF^{c}_{b}, \theta) \tag{4}$$

Where, $AF^{s} \in R^{b\times s}$ indicates that the tensor stores data; s is set to 256 in this article.

Content Feature Extraction Sub-network (CFES). The content presented by the image has a strong correlation with image aesthetics and can further complement the above aesthetic features. Therefore, we designed to use the classic network ResNet-50 [50] as the content feature extraction sub-network. Since unsupervised learning can have a better perception of image content than supervised learning, some work [51] has chosen to use ResNet-50 as the backbone to perform unsupervised learning of image classification tasks on large public data sets. We regard it as an upstream task, learning the content features of image modality information, and then remove the last fully connected layer classifier of ResNet-50 to establish a content feature extraction sub-network, and initialize it using the parameters pre-trained by the upstream task. For the input image, the content feature extraction sub-network performs coding and extraction to obtain a set of content feature tensors $CF^c_{b\times w\times h}$. The content feature tensors can be defined as follows:

$$CF^c_{b\times w\times h} = concat\{cf^1_{b\times w\times h},\ cf^2_{b\times w\times h},\ \ldots,\ cf^c_{b\times w\times h}\} \tag{5}$$

Where, $y^i_{w\times h} \in R^{w\times h}$ represents a certain content feature tensor in the group of content feature tensors, and stores the data in the form of a three-dimensional matrix; b represents batch size; $w \times h$ represents the width and height of the feature tensor, the value is 7×7; c belongs to the channel number, the value is 512; $CF^c_{b\times w\times h} \in R^{b\times c\times w\times h}$ represents that the group of tensors stores data in the form of a four-dimensional matrix.

Similarly, we also change the size and dimension of $CF^c_{b\times w\times h}$ through a feature conversion (FV) module. Specifically, we first mean-pool it to a smaller size and then compress it to a lower dimension, and then adjust the number of channels through a linear layer. The process is defined as follows:

$$CF^c_{b\times 1\times 1} = meanpool(CF^c_{b\times w\times h}) \tag{6}$$

$$CF^c_b = squeeze(CF^c_{1\times 1}) \tag{7}$$

$$CF^s = linear(CF^c_b, \theta) \tag{8}$$

$CF^s \in R^{b\times s}$ represents that the tensor stores data in the form of a two-dimensional matrix; θ represents the parameters in the linear layer; s represents the vector length, which is set to 256 in this article.

Semantic Feature Extraction Sub-network (SFES). Semantic analysis is an advanced way for humans to perceive things. At this level, the human brain further integrates middle-level visual information and associates it with semantic knowledge and context. Perception at this level is related to semantic analysis and conceptual cognition. Specifically, we build a semantic feature extraction sub-network through the pre-trained contextual language reasoning model Bert [49] to extract human high-level perception features, which can well complement the above visual features. For the corresponding comment text of the image, the semantic feature extraction sub-network infers the upper-level semantics with the help of comment context information and provides

feedback with a set of one-dimensional semantic feature tensors SF^c_{length}. The semantic feature tensor can be defined as follows:

$$SF^c_{length} = concat\{sf^1_{length},\ sf^2_{length},\ \ldots,\ sf^c_{length}\} \tag{9}$$

Where, $sf^i_{length} \in R^{b\times length}$ represents a certain semantic feature tensor in the group of semantic feature tensors, and stores data with a two-dimensional matrix; $length$ represents the length of the embedding vector, which is fixed to 768; c belongs to the number of channels, with a value of 128; $SF^c_{length} \in R^{b\times c\times length}$ represents the group of tensors. Quantities store data in the form of three-dimensional matrices.

A feature conversion (FV) module is then applied to change the size and dimension of $CF^c_{b\times w\times h}$. The process is defined as follows:

$$SF^c_{b\times 64} = meanpool(SF^c_{b\times length}) \tag{10}$$

$$SF^d_b = squeeze(SF^c_{b\times 64}) \tag{11}$$

$$SF^s = linear(SF^d_b, \theta) \tag{12}$$

Where, $SF^d_b \in R^{b\times d}$ indicates that the tensor stores data in a two-dimensional matrix, d and s both indicate the vector length, and the sizes are set to 128×64 and 256 respectively.

3.2 Model Training and Prediction Representation

The above three features, CF^s, $AF^s$and SF^s, extracted by the feature extraction subnetwork are spliced and combined to form a multi-modal feature group, denoted as F, which can be defined as follows:

$$F = concat\{CF^s,\ AF^s,\ SF^s\} \tag{13}$$

The extracted multi-modal feature group F is then sent to Multi-Layer Perceptron (MLP) for further prediction and output aesthetic distribution d. The process can be described as follows:

$$d = MLP(F, \theta) \tag{14}$$

Where, θ represents the trainable parameter in MLP. We set up two hidden layers, with the number of neurons being 256 and 128, respectively. The PRelu activation function is used between the hidden layers. Since the number of neurons is relatively small, we do not use Dropout.

The predicted image aesthetic score is obtained based on the aesthetic distribution $\hat{Y}$. The calculation method is as follows:

$$\hat{Y} = \sum_{i=1}^{10} d_i p_i \tag{15}$$

Where, d_i represents the i-th discrete point in aesthetic distribution d, and p_i represents the probability value of this discrete point.

Following the previous works [24], [39], We use the EMD loss function $L_{EMD}(d, \hat{d})$ to assist in optimizing our model, which is defined as follows:

$$L_{EMD}(d, \hat{d}) = (\frac{1}{N}\sum_{k=1}^{N}|CDF_d(k) - CDF_{\hat{d}}(k)|^r)^{1/r} \quad (16)$$

Where, $CDF_d(k)$ represents the probability value of a certain point in the predicted aesthetic distribution.$CDF_{\hat{d}}(k)$ represents the probability value label value of a certain point in the aesthetic distribution.Nrepresents the number of aesthetic score values in the aesthetic distribution, and here it sets to 10, that is, 1-10 points.$r$takes a value of 2 to speed up gradient convergence.

4 Experimental Results

We first introduce our experimental setup to demonstrate the performance of our model in this section. The superiority of our model is proved by extensive experiments on public image aesthetic assessment databases with some image aesthetic assessment models, and finally some ablation experiments is conducted to demonstrate the effectiveness of each sub-network through.

4.1 Experimental Setups

Implementation Details. For all input images I, we resize them to 224×224×3, and for each review text, we convert it into a token through a pre-trained tokenizer as the input of the BERT model [49]. The Adam optimizer with an initial learning rate of 3×10^{-5}, and the decay rate of 0.9 times the original value after every 5 epochs. The batch size is set to 40. All experiments were implemented under the PyTorch framework and accelerated training was performed on a computer with a single NVIDIA GeForce RTX 3090 24G GPU.

Evaluation Criteria. We use several commonly used evaluation metrics to evaluate the prediction accuracy of the model: accuracy (ACC), Spearman's rank correlation coefficient (SRCC) [24] and Pearson's linear correlation coefficient (PLCC) [21]. Earth Mover Distance (EMD) was utilized to evaluate the performance of aesthetic distribution predictions. SRCC is calculated as follows:

$$SRCC = 1 - \frac{6\sum_{i=1}^{N}(d_i)^2}{N(N^2-1)} \quad (17)$$

Where d_i represents the difference between the real score and the predicted score; N represents the number of images.

PLCC is calculated as follows:

$$LCC = \frac{\sum_{i=1}^{N}(y_i-\bar{y})(\hat{y}_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2}} \quad (18)$$

Where N is the number of images; y_i and $\hat{y}_i$ represents the real aesthetic value and predicted aesthetic score of the i-th image respectively; $\bar{y}$ and $\hat{y}$ represents the real aesthetic average and predicted average value respectively.

ACC is calculated as follows:

$$Acc = \frac{TP+TN}{TP+FN+FP+TN} \tag{19}$$

Where TP represent the positive prediction, and the actual value is positive; FP means the positive prediction, and the actual value is negative; TN means the negative prediction, and the actual value is positive; FN means the negative prediction, and the actual value is negative.

4.2 Databases

AVA database [53]. All images in this database are collected from the DPChallenge website. The total number of images has exceeded 250,000. The database covers various common types of images. Following previous research, the images are included training set with 229,951 images, validation set with 12,775 images, and test set with 12,776 images.

AVA-Comments database [54]. The AVA-Comments database crawls the user comments corresponding to all images in the AVA data set on the DPChallenge website and obtains more than 1.5 million user comments. It removes all quotation marks and additional HTML tags in the crawled data and converts the data into Concentrate the valid comments of each image into a single text file as the text modal data of the corresponding image.

4.3 Comparison with Advanced Methods

We applied 6 IAA models to conduct the fair comparison on the AVA database, and the experimental results are summarized in Table 1. Since some models are released without source code or related data in the paper, these results are marked with "-". The best performance for each metric is marked in bold. Overall, our model achieves superior performance. Specifically, NIMA [24] was a classic model for early image aesthetic assessment with simple structure and didn't yield a good result. HLA-GCN [39] used a novel graph convolution method to extract image features, which has improved the accuracy compared to work [24]. MLSP [40], TANet [41] and TAVAR [42] all considered multi-level features of the image, so these models achieved further improvement on SRCC and PLCC indicators. It needs to be mentioned that the above works are all built based on a single modality. Multimodal co-TRM [47] took into account the aesthetic features and text features of multimodal information and have the best classification accuracy among all the models, which also proves the reasonable of image and text feature fusion. Our MFD model not only considers multi-modal information, but also adds additional content features at the visual level as a supplement, which yields best

SRCC and PLCC indicators, and the model's classification accuracy also achieves competitive results.

Table 1. Comparison of the Spearman rank correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC) and classification accuracy (ACC) between the proposed model and other image aesthetics assessment models

Method	backbone	SRCC	PLCC	ACC	EMD
NIMA [24]	Inception-v2	0.612	0.636	81.51	0.050
HLA-GCN [39]	ResNet-50	0.665	0.687	84.60	0.043
MLSP [40]	Inception-ResNet	0.756	0.757	81.72	-
TANet [41]	MobileNet-v2	0.753	0.762	80.01	0.049
TAVAR [42]	ResNet-50	0.725	0.736	85.10	-
Multimodal co-TRM [47]	FCN	0.784	-	**85.63**	-
MFD(Ours)	Efficient-net_v2_s	**0.800**	**0.812**	82.49	**0.038**

4.4 Ablation Studies

In order to verify the effectiveness of each sub-network of the proposed model, in this section, we deeply explore the impact of each sub-network of the model on the overall performance of the model. All ablation studies are implemented based on the AVA database [53] and AVA-Comments database [10]. We divide the model as a whole into Aesthetic Feature Extraction Sub-network (AFES), Content Feature Extraction Sub-network (CFES) and Semantic Feature Extraction Sub-network (SFES). Table 2 shows the evaluation indicators under different sub-network combinations. The best performance for each metric is marked in bold. Experimental results show that when each sub-network is executed individually, the integrity of the model is low, and the performance of AFES is better than that of other sub-networks, which further illustrates that aesthetic features play an important role in image aesthetic assessment. When AFES and SFES are combined with CFES and SFES, the overall performance is good, while with only AFES and CFES, the model performance is average. This shows that different modal information can provide the model with richer feature representations, thereby improving the overall performance of the model. Compared with the combination of AFES, CFES and SFES, the overall performance of the combination of AFES and SFES is relatively close. Our subjective analysis is due to the lack of expression ability of the content features extracted by CFES. Limited to upstream tasks, we are temporarily unable to explore CFES of other backbones, and we will further conduct related work in the future. Finally, it can be obtained that the combination of three sub-

networks achieves the best model performance and prediction accuracy, which also proves the effectiveness of our proposed MFD method.

Table 2. Comparison of the Spearman rank correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC) and classification accuracy (ACC) of the proposed model for ablation studies of different feature extraction sub-networks

Method	backbone	SRCC	PLCC	ACC	EMD
AFES	Efficient-net_v2_s	0.704	0.716	80.60	0.045
CFES	ResNet-50	0.643	0.655	78.46	0.052
SFES	Bert	0.642	0.664	75.71	0.052
AFES+CFES	Efficient-net_v2_s + ResNet-50	0.708	0.719	80.42	0.045
AFES+SFES	Efficient-net_v2_s + Bert	0.795	0.808	82.45	0.038
CFES+SFES	ResNet-50+ Bert	0.770	0.785	81.54	0.040
AFES+CFES +SFES	Efficient-net_v2_s + ResNet-50+ Bert	**0.800**	**0.812**	**82.49**	**0.038**

4.5 Performance of Different Backbones

Considering that different backbones have different feature representation capabilities, and the aesthetic features play a vital role in image aesthetic assessment, in this part, we explore the specific impact of different backbones in AFES on the model. The specific performance of each model can be seen in Table 3. The best performance for each metric is marked in bold.

It can be seen that when AFES uses Efficientnet_v2_s as the backbone, the model evaluation index is greatly improved compared to other backbones, with a maximum improvement of 0.034 for SRCC, a maximum improvement of 0.027 for PLCC, and a maximum improvement of ACC of 1.41%. The performance of the model is relatively good when using Densenet121 [55], but it is still slightly lower than Efficientnet_v2_s in various evaluation indicators. In general, the backbones can slightly affect the overall performance, and the specific task can choose a suitable model for the task.

Table 3. Comparison of the Spearman rank correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC) and classification accuracy (ACC) of the proposed model using different backbones on AFES

Method	Backbone (AFES)	SRCC	PLCC	ACC	EMD
AFES+ CFES (ResNet-50) + SFES (Bert)	AlexNet [36]	0.766	0.785	81.08	0.040
	VGG16 [12]	0.773	0.790	81.47	0.039
	ResNet-18 [50]	0.770	0.787	81.47	0.039
	ResNet-50 [50]	0.774	0.789	81.89	0.039
	Googlenet [56]	0.773	0.790	81.32	0.039
	Densenet121 [57]	0.780	0.796	82.04	0.038
	Efficient-net_v2_s [52]	0.800	0.812	82.49	0.038

5 Conclusion

In this paper, we propose an image aesthetic assessment method (MFD) driven by multi-modal features from the perspective of different levels of object analysis by the human brain. For different modal information, a specific method of feature extraction sub-network is designed to extract dissimilar categories of features corresponding to the modal content. Specifically, for image content, the visual perception of the human eye dominates, so we design the Aesthetic Feature Extraction Sub-network (AFES) and Content Feature Extraction Sub-network (CFES) to extract image aesthetic features and image content features, respectively. For text content, high-level semantic understanding is the main analysis method of the human brain, and a Semantic Feature Extraction Sub-network (SFES) is designed to extract text semantic features. In order to fuse multiple types of features, we designed a feature conversion (FV) module to change the dimensions and sizes of different features. The obtained three types of features are then used to predict the aesthetic score distribution.

Through experiments on a public image aesthetic assessment dataset, the effectiveness and superiority of our proposed model are proved based compared with mainstream image aesthetic assessment methods.

Acknowledgments. This work is supported by Liaoning Province Natural Science Foundation under Grant 2023-MS-139, Shenyang science and technology plan project under Grant 23-407-3-32 and National Natural Science Foundation of China under Grant 61901205.

References

1. Li, L., Zhu, T., Chen, P., Yang, Y., Li, Y., & Lin, W.: Image Aesthetics Assessment with Attribute-Assisted Multimodal Memory Network. In: IEEE Transactions on Circuits and Systems for Video Technology, pp. 1-1 (2023)
2. Yan, B.: A CNN-LSTM-based model for fashion image aesthetic captioning. In: Proceedings of SPIE - The International Society for Optical Engineering, 12511 (2023)
3. Silva, W., Carvalho, M., Mavioso, C., Cardoso, M.J., Cardoso, J.S.: Deep Aesthetic Assessment and Retrieval of Breast Cancer Treatment Outcomes. In: Pinho, A.J., Georgieva, P., Teixeira, L.F., Sánchez, J.A. (eds) Pattern Recognition and Image Analysis. IbPRIA 2022. Lecture Notes in Computer Science, vol 13256. Springer, Cham (2022)
4. Wang, L., Wang, X., & Yamasaki, T.: Image aesthetics prediction using multiple patches preserving the original aspect ratio of contents. In: Multimed Tools Appl 82, pp. 2783–2804 (2023)
5. Pandit, A., Animesh, Gautam, B.K., Agarwal, R.: Image Aesthetic Score Prediction Using Image Captioning. In: Kumar, A., Mozar, S., Haase, J. (eds) Advances in Cognitive Science and Communications. ICCCE 2023. Cognitive Science and Technology. Springer, Singapore (2023)
6. Ke, J., Ye, K., Yu, J., Wu, Y., Milanfar, P., & Yang, F.: VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2023-June, pp. 10041–10051 (2023)
7. Naderi, M.R., Givkashi, M.H., Karimi, N. et al.: Aesthetic-aware image retargeting based on foreground–background separation and PSO optimization. In: Multimed Tools Appl (2023)
8. Zhang, X., Gao, X., He, L., & Lu, W.: MSCAN: Multimodal Self-and-Collaborative Attention Network for image aesthetic prediction tasks. In: Neurocomputing, 430, pp. 14–23 (2021)
9. Y. Cui, G. Jiang, M. Yu, et al.: Stitched wide field of view light field image quality assessment: benchmark database and objective metric. In: IEEE Transactions on Multimedia, early access, 2023, doi: 10.1109/TMM.2023.3330096 (2023)
10. Krizhevsky, A., Sutskever, I., & Hinton, G. E.: ImageNet classification with deep convolutional neural networks. In: Communications of the ACM, **60**(6), pp. 84–90 (2017)
11. Kao, Y., Wang, C., & Huang, K.: Visual aesthetic quality assessment with a regression model. In: Proceedings - International Conference on Image Processing, ICIP, 2015-December pp. 1583–1587 (2015)
12. Simonyan, K., & Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (2015)
13. Mai, L., Jin, H., & Liu, F.: Composition-preserving deep photo aesthetics assessment. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, pp. 497–506 (2016)
14. Zeng, H., Zhang, L., & Bovik, A. C.: A probabilistic quality representation approach to deep blind image quality prediction. In: arXiv (2017)

15. Liu, L., Guo, X., Bai, R., & Li, W.: Image Aesthetic Assessment Based on Attention Mechanisms and Holistic Nested Edge Detection. In: Proceedings - 2022 Asia Conference on Advanced Robotics, Automation, and Control Engineering, ARACE 2022, pp. 70–75 (2022)
16. Black, K., Janner, M., Du, Y., Kostrikov, I., & Levine, S.: Training Diffusion Models with Reinforcement Learning. In: arXiv (2023)
17. Valenzise, G., Kang, C., Dufaux, F.: Advances and Challenges in Computational Image Aesthetics. In: Ionescu, B., Bainbridge, W.A., Murray, N. (eds) Human Perception of Visual Information. Springer, Cham (2022)
18. Biswas, K., Shivakumara, P., Pal, U. et al.: Classification of aesthetic natural scene images using statistical and semantic features. In: Multimed Tools Appl 82, pp. 13507–13532 (2023)
19. Jang, H., Lee, Y., & Lee, J.-S.: Modeling, Quantifying, and Predicting Subjectivity of Image Aesthetics. In: arXiv (2022)
20. Zhu, T., Li, L., Chen, P., Wu, J., Yang, Y., Li, Y., & Guo, Y.: Attribute-assisted Multimodal Network for Image Aesthetics Assessment. In: Proceedings - IEEE International Conference on Multimedia and Expo, 2023-July, pp. 2477–2482 (2023)
21. Withöft, A., Abdenebaoui, L., Boll, S.: ILMICA - Interactive Learning Model of Image Collage Assessment: A Transfer Learning Approach for Aesthetic Principles. In: Þór Jónsson, B., et al. MultiMedia Modeling. MMM 2022. Lecture Notes in Computer Science, vol 13142. Springer, Cham (2022)
22. Ramachandram, D., Taylor, G. W.: Deep multimodal learning: A survey on recent advances and trends. In: IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 96–108 (2017)
23. Zhu, W., Wang, X., Li, H.: Multi-modal deep analysis for multimedia. In: IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 10, pp. 3740–3764 (2019)
24. Talebi, H., Milanfar, P.: Nima: Neural image assessment. In: IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 3998–4011 (2018)
25. Datta, R., Joshi, D., Li, J., Wang, J. Z.: Studying aesthetics in photographic images using a computational approach. In: Proceedings of the European Conference on Computer Vision. Springer, 2006, pp. 288–301 (2006)
26. Wong, L.-K., Low, K.-L.: Saliency-enhanced image aesthetics class prediction. In: Proceedings of the IEEE International Conference on Image Processing. IEEE, 2009, pp. 997–1000 (2009)
27. Xue, W., Zhang, L., Mou, X.: Learning without human scores for blind image quality assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 995–1002. 1 (2013)
28. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1733–1740. 1 (2014)
29. Bosse, S., Maniry, D., Wiegand, T., Samek, W.: A deep neural network for image quality assessment. In: Image Processing (ICIP), 2016 IEEE International Conference on. IEEE, 2016, pp. 3773–3777. 1 (2016)
30. Bianco, S., Celona, L., Napoletano, P., Schettini, R.: On the use of deep learning for blind image quality assessment. In: arXiv preprint arXiv:1602.05531, 2016. 1, 6, 8 (2016)
31. Lu, X., Lin, Z., Shen, X., Mech, R., Wang, J. Z.: Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 990–998. 1 (2015)
32. Kao, Y., Huang, K., Maybank, S.: Hierarchical aesthetic quality assessment using deep convolutional neural networks. In: Signal Processing: Image Communication, vol. 47, pp. 500–510 (2016)

33. Mai, L., Jin, H., Liu, F.: Composition-preserving deep photo aesthetics assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 497–506 (2016)
34. Zeng, H., Zhang, L., Bovik, A. C.: A probabilistic quality representation approach to deep blind image quality prediction. In: arXiv preprintarXiv:1708.08190, 2017. 1 (2017)
35. Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking network with attributes and content adaptation. In: European Conference on Computer Vision. Springer, 2016, pp. 662–679. 1, 6, 7 (2016)
36. Ma, S., Liu, J., Chen, C. W.: A-lamp: Adaptive layout-aware multipatch deep convolutional neural network for photo aesthetic assessment. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017. 1, 6, 7 (2017)
37. Li, L., Zhu, H., Zhao, S., Ding, G., Lin, W.: Personality-assisted multitask learning for generic and personalized image aesthetics assessment. In: Proceedings of IEEE Transactions on Image Processing, vol. 29, pp. 3898–3910 (2020)
38. Li, L., Zhi, T., Shi, G., Yang, Y., Xu, L., Li, Y., Guo, Y.: Anchor-based knowledge embedding for image aesthetics assessment. In: Proceedings of NEUROCOMPUTING,2023 (2023)
39. She, D., Lai, Y. K., Yi, G., et al.: Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: pp. 8475-8484 (2021)
40. Hosu, V., Goldlucke, B., Saupe, D.: Effective aesthetics prediction with multi-level spatially pooled features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: pp. 9375-9383 (2019)
41. He, S., Zhang, Y., Xie, R., et al.: Rethinking image aesthetics assessment: Models, datasets, and benchmarks. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. 2022: pp. 942-948 (2022)
42. Li, L., Huang, Y., Wu, J., et al.: Theme-aware Visual Attribute Reasoning for Image Aesthetics Assessment. In: IEEE Transactions on Circuits and Systems for Video Technology, 2023 (2023)
43. Zhang, R., Zhang, Z., Li, M., Ma, W.-Y., Zhang, H.-J.: A probabilistic semantic model for image annotation and multimodal image retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 1. IEEE, 2005, pp. 846–851 (2005)
44. Wu, Q., Wang, Z., Deng, F., Chi, Z., Feng, D. D.: Realistic human action recognition with multimodal feature selection and fusion. In: IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 43, no. 4, pp. 875–885, 2013 (2013)
45. He, X., Peng, Y.: Fine-grained image classification via combining vision and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5994–6002 (2017)
46. Zhang, X., Gao, X., Lu, W., He, L., Li, J.: Beyond Vision: A Multimodal Recurrent Attention Convolutional Neural Network for Unified Image Aesthetic Prediction Tasks. In: IEEE Transactions on Multimedia, vol. 23, 2021, pp. 611-623 (2021)
47. Miao, H., Zhang, Y., Wang, D., Feng, S.: Multimodal Aesthetic Analysis Assisted by Styles through a Multimodal co-transformer Model. In: Proceedings of the IEEE 24th International Conference on Computational Science and Engineering (CSE), 2021 (2021)
48. Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H., Qian, J.: Multimodal sentiment analysis with image-text interaction network. In: IEEE Transactions on Multimedia, pp. 1-12, 2022 (2022)
49. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019 - 2019 Conference

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 4171–4186 (2019)
50. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: pp. 770-778 (2016)
51. He, K., Fan, H., Wu, Y., et al.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: pp. 9729-9738 (2020)
52. Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: International conference on machine learning. PMLR, 2021: pp. 10096-10106 (2021)
53. 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. IEEE, 2012: pp. 2408-2415 (2012)
54. Zhou, Y., Lu, X., Zhang, J., et al.: Joint image and text representation for aesthetics analysis. In: Proceedings of the 24th ACM international conference on Multimedia. 2016: pp. 262-266 (2016)
55. Huang, G., Liu, Z., Van Der Maaten, L., et al.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: pp. 4700-4708 (2017)
56. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: arXiv preprint arXiv:1409.1556, 2014 (2014)
57. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: pp. 1-9 (2015)