mathematics

Article

Multimodal Image Aesthetic Prediction with Missing Modality

Xiaodan Zhang [1,†], Qiao Song [1,†] and Gang Liu [2,*]

[1] Science and Technology of Information Institute, Northwest University, Xi'an 710127, China; xiaodanzhang@nwu.edu.cn (X.Z.); songqiao@stumail.nwu.edu.cn (Q.S.)
[2] Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China
* Correspondence: liugang@opt.cn
† These authors contributed equally to this work.

Abstract: With the increasing growth of multimedia data on the Internet, multimodal image aesthetic assessment has attracted a great deal of attention in the image processing community. However, traditional multimodal methods often have the following two problems: (1) Existing multimodal image aesthetic methods are based on the assumption that full modalities are available in all samples, which is unapplicable in most cases since textual information is more difficult to obtain. (2) They only fuse multimodal information at a single level and ignore their interaction at different levels. To address these two challenges, we proposed a novel framework termed Missing-Modility-Multimodal-Bert networks (MMMB). To achieve the completeness, we first generate the missing textual modality conditioned on the available visual modality. We then project the image features to the token space of the text, and use the transformer's self-attention mechanism to make the two different modalities information interact at different levels for earlier and more fine-grained fusion, rather than only at the final layer. A large number of experiments on two large benchmark datasets in the field of image aesthetic quality evaluation: AVA and Photo.net demonstrate that the proposed model significantly improves image aesthetic assessment performance under both textual missing modality condition and full-modality condition.

Keywords: image aesthetic quality assessment; multimodal learning; missing multimodal data; transformer

MSC: 68T05

check for updates

Citation: Zhang, X.; Song, Q.; Liu, G. Multimodal Image Aesthetic Prediction with Missing Modality . *Mathematics* **2022**, *10*, 2312. https://doi.org/10.3390/math10132312

Academic Editor: Junzo Watada

Received: 31 May 2022
Accepted: 29 June 2022
Published: 1 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

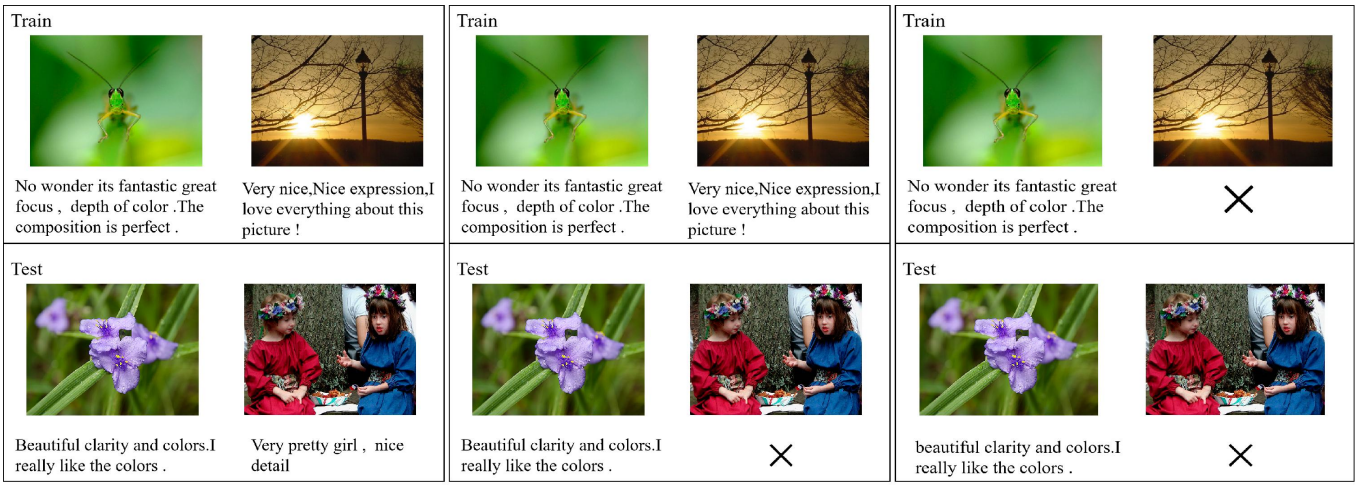
Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Aesthetics, in the world of photography, refers to the appreciation of beauty in the form of art. Image aesthetic quality assessment aims to use computers to simulate human perception of aesthetics and automatically evaluate the aesthetics of images. It has found great applications in many areas, such as photo ranking [1], image recommendation [2], and image retrieval and editing [3]. Thus, image aesthetic assessment has attracted increasing attention in recent years [4–9].

In the early stages, the research of image aesthetics mainly focuses on designing hand-crafted features according to the photographic rules such as the lightning, color, and global image layout. Such methods first extract hand-crafted features from images and then learn a mapping of these features to subjective aesthetic quality [4–6]. Later, with the proposal and development of convolution neural network [10,11], deep features have been used to capture the low-level and high-level descriptive aesthetic attributes, which significantly improves the performance of image aesthetic quality evaluation tasks [7–9]. However, most of these methods are adapted from classical image classification networks, not specific to image aesthetic quality assessment tasks and often focus only on image features without considering other relevant data resources, thus the performance is limited.

With the popularity of the Internet, the modern digital world we live in is multimodal in nature: images on the web are often accompanied with text. For example, on Photo.net (https://www.photo.net/, accessed on 1 May 2022), Dpchallenge (https://www.dpchallenge.com/, accessed on 1 May 2022) and other image sharing websites, users are allowed to make subjective comments on images. The comments describe the content of the image and the feelings it brings to people; thus, it is helpful in image aesthetic assessment tasks. Recently, a mushrooming number of works have been proposed to exploit the textual features. For example, Zhou et al. [12] utilize multimodal Deep Boltzmann Machine (DBM) to encode both visual and textual information. Zhang et al. [13] propose a Multimodal Recurrent Attention Convolutional Neural Network (MRACNN) to leverage the semantic power of user comments. Although the aforementioned methods obtain effective results, they often assume the completeness of modality in both training and test data as illustrated in Figure 1. However, such an assumption may not always hold in the real world. For example, we may not be able to access textual data since many voters only give aesthetic scores without textual comments. Thus, an interesting yet challenging research question then arises: Can we learn a multimodal image aesthetic quality assessment model from an incomplete dataset while its performance should as close as possible to the one that learns from a full-modality dataset? In addition, existing multimodal methods [12,14] often use the concatenation or element-wise summations for multimodal feature fusion. Since the visual features and the textual features may vary significantly, traditional multimodal fusion methods may be insufficient, limiting the final prediction performance.



(a) Train:full modality Test:full modality

(b) Train:full-modality Test: missing modality

(c) Train: missing modality Test: missing modality

Figure 1. Motivation of our approach. (**a**) Train and test with full modality (Zhang et al. [13,14]); (**b**) testing with missing modality (Zhou et al. [12]); (**c**) we study missing textual modality in training, testing, or both.

In this paper, we systematically study this problem and propose the Missing-Modility-Multimodal-Bert (MMMB) model. In order to deal with the missing textual modality at any stage, we reconstruct the missing textual description related to aesthetics according to the available image information. The generated textual modality along with the visual features are sent into the token spaces. Then, we make full use of the multi-head self-attention in the Bert model to fuse the multimodalities information at different levels, rather than only at the final layer.

The contributions of this paper are as follows:

- To the best of our knowledge, we are the first work to systematically study the problem of missing modality data in both train set and test set in the field of multimodal image aesthetic quality assessment.
- Inspired by self-supervised pretraining of Transformer-style architectures, we use the multi-head self-attention mechanism to capture the complex associations between the visual features from images and the textual features from comments at different levels. The experimental results demonstrate that the proposed multimodal fusion module significantly outperforms existing multimodal approaches.
- We conducted extensive experiments to prove the superiority of the proposed method over other latest works on a two benchmark dataset.

The remainder of this paper is organized as follows: Section 2 summarizes the related work. Section 3 introduces the proposed approach. Section 4 quantitatively analyses the effectiveness of the proposed method and compares it with state-of-the-art results. Finally, Section 5 contains a summary and plans for future work.

2. Related Work

In this section, we briefly reviewed the related work on the following two research topics: (1) image aesthetic assessment and (2) missing modality problem.

2.1. Image Aesthetic Assessment

Unimodal learning. The task of image aesthetic quality assessment begins with only attention to image information, that is, unimodal learning. How to distinguish between photos taken by professional photographers and photos taken by ordinary users [15] is the earliest attempt of researchers in the field of image aesthetic quality assessment. The traditional method is based on the aesthetic features of artificial design [16,17]. Most of these handcrafted features are inspired by photography rules, such as clarity, depth of field, colorfulness, rule of third, etc. Although these methods achieved good performance at that time, they could not accurately capture complex aesthetic factors and had certain limitations. With the development of convolutional neural network, some methods based on deep learning have been proposed, which greatly improves the performance of image aesthetic assessment. Lu et al. [3] first tried to apply convolutional neural networks to this field and proposed a double-column deep convolutional neural network architecture to learn the global and local features of the image, respectively, and finally complete the aesthetic binary classification task. Kao et al. [18] used a regression model instead of a common classification model to evaluate image aesthetics. They believe that continuous scores can better express the aesthetic quality of images. Talebi et al. [19] use Earth Mover's Distance (EMD) loss, which is different from the previous loss function, to optimize the network, and propose a new task to predict the distribution of aesthetic scores. The lack of aspect ratio information in an image will affect the predicted aesthetic score. Wang et al. [20] proposed a multi-patch training method that maintains the aspect ratio to predict the aesthetic score of images. However, most of these methods are often focused only on image features without considering other relevant data sources and higher-level semantic information, thus the performance is limited.

Multimodal learning. Multimodal learning uses complementary information from different modalities to improve the performance of various computer vision tasks [21]. With the release of various excellent models in the field of natural language processing, the text comment information of images in the field of image aesthetic quality assessment has also attracted the attention of researchers. In multimodal image aesthetic prediction tasks, how to effectively fuse the information of each modality is a key point. Zhou et al. [12] introduced multimodal learning into the field of image aesthetic quality assessment for the first time. In addition to image information, they also paid attention to higher-level semantic information. They used DBM to jointly represent image information and text information for aesthetic assessment, and built the AVA-Comments dataset for researchers to use. Hii et al. [22] use the MultiGap deep network architecture to extract image features and

use RNN to extract text features. Finally, the visual and textual features are concatenated directly to predict the aesthetic distribution of the input image. With the proposal of attention mechanism, researchers try to apply it to the task of image aesthetic quality assessment. Instead of focusing on each part of the information as before, they selectively pay attention to some key areas. Zhang et al. [14] employ a recurrent attention network, which can eliminate irrelevant information in images and extract visual features only in important regions. In terms of multimodal fusion, they use MFB to model the correlation between different multimodal features, which has achieved good performance. In another work, Zhang et al. [13] used self-attention to encode the interdependence between visual elements in images when extracting visual features, so as to effectively capture the global composition of images. Co-attention is used to capture the intrinsic correlation between two modalities for more effective feature fusion. Miao et al. [23] propose an end-to-end multi-output deep learning model based on multimodal GCN and co-attention for image aesthetics and emotion conjoint analysis. Although the aforementioned methods obtain effective results, they use late fusion manner to fuse multimodal inputs. This late fusion layer usually needs multimodal data to exist at the same time in the training stage. However, for multimodal image aesthetic tasks, acquiring enough textual modality data is still very challenging and expensive.

2.2. Missing Modality Problem

Recently, some methods have been proposed to solve the problem of modality missing in multimodal learning. Tran et al. [24] used a cascaded residual autoencoder for imputation of missing modalities. It consists of a set of stacked residual autoencoders, which iteratively simulate the residual between the current prediction and the original data. Ma et al. [25] proposed the use of Bayesian meta-learning framework to reconstruct missing modalities and regularize the reconstructed latent features, effectively dealing with the problem of missing modalities. Zhang et al. [26] aim to use feature reconstruction to learn a joint multimodal representation of the latent space that can contain relevant information from all modalities, for supervised learning of predicting the target. The above methods can solve the problem of missing modality data to a certain extent, but the training process is relatively cumbersome. It is not applicable to the large scale datasets used in the field of image aesthetic quality assessment. In this paper, we adopt a more direct and concise method, that is, generating the textual description related to aesthetics according to the available image information when the textual modality is missing.

3. Method

Problem Formulation. In multimodality image aesthetic prediction problems, we are given a multimodal dataset containing two modalities, i.e., image and text comments. Formally, we let $D = \{D^f, D^m\}$ denote a multimodal image aesthetic dataset. $D^f = \{x_i^1, x_i^2, y_i\}$ represent the full-modality samples, where x_i^1 and x_i^2 represent visual modality and textual modality of i-th sample respectively, and y_i is the corresponding aesthetic label. $D^m = \{x_i^1, y_i\}$ are the modality-incomplete samples. We assume visual modality is available for all samples, while textual modality is available for only a portion of the subjects. Our target is to leverage both modality-complete and modality-incomplete data for model training.

In this paper, we propose a novel multimodal image aesthetic quality assessment method, i.e., MMMB model. The overview of the approach is shown in Figure 2. For the full-modality samples, we first use the image encoder to extract the raw features of the image and then map the extracted visual features to the token space. In the token spaces, the visual features, connected with the textual embedding features, are sent to the multimodal encoder, which uses the self-attention mechanism for multi-level and fine-grained fusion. In the case of missing textual modality, we generated the missing textual modality conditioned on existing modality images and then form a multimodal

joint representation. Finally, the feature vectors output by the multimodal encoder are sent to the aesthetic prediction layer for aesthetic value assessment.

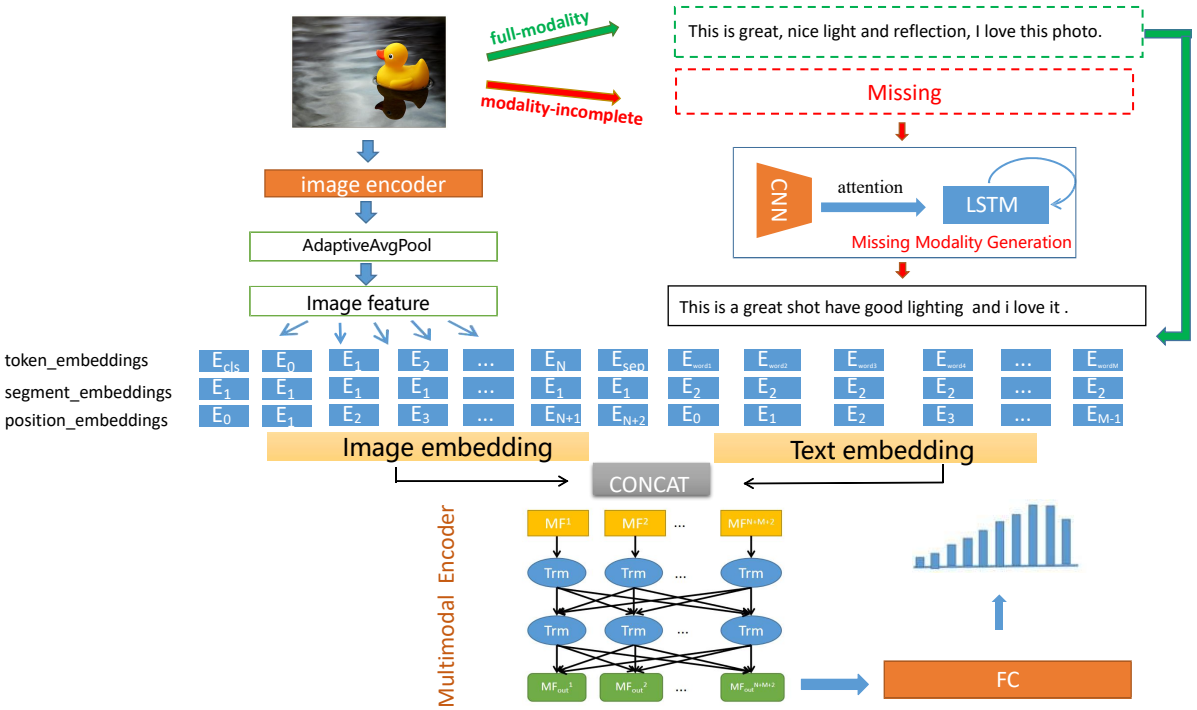


Figure 2. The framework of the proposed method. For full-modality samples, the processing process is shown in the green arrow, and for modality-incomplete, the processing process is shown in the red arrow. The image encoder is used to extract image features, and the missing modality generation network is used to generate the missing textual aesthetic description when the input is modality-incomplete. The embedding module generates the input of the multimodal encoder. The multi-layer self-attention mechanism is used to make the information of two different modalities interact at different levels for earlier and more fine-grained feature fusion.

3.1. Image Encoder

Experiments have proved that the convolutional neural network architecture pre-trained on ImageNet [10] can be used to extract more effective visual features. In our method, any type of CNN can be used as a visual feature extractor. We take Resnet50 [27] as an example in this section. The input image I is first resized to 224×224 and then fed into the CNN to extract the deep features. In order to obtain N independent embeddings consistent with the text information, we replace the original pool layer with an Adaptive-pool layer. In addition, the output of the feature map is thus a tensor with dimensions (W, H, D), where W and H represent the spatial resolution, and D is the channel dimension. The extractor produces $N = W * H$ vectors, which can be represented as follows:

$$f(I, \quad i) = \left\{ r_i \mid r_i \in R^D, i = 1, 2, \ldots, N \right\} \tag{1}$$

where $f(\cdot, i)$ represents D-dimensional representation corresponding to a part of the image.

3.2. Missing Modality Generation

For the problem of missing modality data, traditional methods often directly discard the samples with missing modality data or reconstruct a multimodal joint representation of hidden space to encode multimodal information. However, these methods either lead to the reduction of sample count and loss of some important information, or need to update all samples at the same time, which is not applicable on large scale datasets for the image aesthetic quality assessment task. In this paper, we generate feature representation of

missing textual modality in the latent space based on the available visual modality. Given the observed visual modality x_1, in order to obtain the reconstruction x_2 of the missing modality, we optimize the following objective for the reconstruction network:

$$\theta^* = \arg\max_{\theta} \sum_{\{x_1, x_2\}} -\log p(x_2 \mid x_1; \theta) \tag{2}$$

where θ are the parameters of our reconstruction model. However, it is non-trivial to train a reconstruction network from limited modality complete samples. Inspired by [28], we use an attention-based approach to generate approximate feature representations of textual modality using LSTM networks by attention to the salient part of an image. In order to learn the general aesthetic textual representation and reduce the complexity of the network, we first pretrained the reconstruction network on the DPC dataset [29] (DPC-Captions dataset is an image aesthetic caption dataset, which contains 154,384 images and 2,427,483 comments from DPChallenge.com (accessed on 1 May 2022)).

Specifically, given the observable visual modality, the convolution neural network is used to extract the visual features $x = \{x_1, x_2 \ldots, x_L \mid x_i \in R^D\}$. The attention weight α_i, which is used to measure the weight of the image feature at the i-th position when generating the *t-th* word, is calculated for each position i:

$$e_{ti} = f_{att}(x_i, h_{t-1}) \tag{3}$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})} \tag{4}$$

where f_{att} is a multilayer perception, h_{t-1} is the previous hidden state. We can calculate the context vector $\hat{z}_t$ after obtaining the attention weight:

$$\hat{z}_t = \psi(\{x_i\}, \{\alpha_i\}) = \sum_{i}^{L} \alpha_{ti} x_i \tag{5}$$

Then, a long short-term memory (LSTM) [30] network is used to produce the missing textual comments conditioned on a context vector $\hat{z}_t$, the previous hidden state h_{t-1}, and the previously generated words. Our implementation of LSTM follows the one used in [31]:

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ g_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} T_{D+m+n,n} \begin{pmatrix} Ey_{t-1} \\ h_{t-1} \\ \dot{z}_t \end{pmatrix} \tag{6}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ g_t \tag{7}$$

$$h_t = o_t \circ \tanh(c_t) \tag{8}$$

where i_t, f_t, c_t, o_t, and h_t are the input, forget, memory, output, and hidden state of the LSTM, respectively. $E \in R^{m \times k}$ is an embedding matrix, σ represents the sigmoid activation function, and $\circ$ represents the element-wise multiplication.

By using a deep output layer [32], we can calculate the probability of each word in the wordmap as follows:

$$p\left(y_t \mid x, y^{t-1}\right) \propto \exp(L_o(Ey_{t-1} + L_h h_t + L_z \hat{z}_t)) \tag{9}$$

where $L_o \in R^{k \times m}$, $L_h \in R^{m \times n}$,$L_z \in R^{m \times D}$ and E are learned parameters initialized randomly. Finally, the word with the highest probability is taken as the currently generated word and used as the input of the next time.

3.3. Embedding Model

In this section, we will introduce the Embedding Layer, which generates input for the multimodal encoder.

3.3.1. Segment and Position Embedding

Segment embedding is used to distinguish different modalities. Specifically, we assign a segment ID to image modality and text modality, respectively. In specific experiments, we set the segment ID of image modality to 0and the segment ID of text modality to 1. Position embedding represents each embedding the relative position information in the segment. Each segment is counted from 0.

3.3.2. Text Embedding

For full-modality samples, we input the text comment rounds in the comment dataset. However, for the samples with missing textual modality, we input the reconstructed textual modality data. We adopt the same coding method as Bert [33] to process text input, which firstly divides text into a word sequence, and then uses the WordPiece [34] method to tokenize each word. Then, the token embedding is transformed into a 768-dimensional vector representation. We use $t = \{t_1, t_2, \ldots, t_M\} \in R^d$ to represent the input text sequence, where M represents the number of words, and d represents the embedded 768 dimension. Similar to the traditional Bert, we add position embedding and segment embedding. The final text comment can be represented as $\{\hat{t}_1, \hat{t}_2, \ldots \hat{t}_M\}$, and the textual representation at the i-th position is calculated by:

$$\hat{t}_i = LayerNorm(t_i + p_i + s_t) \tag{10}$$

where LayerNorm [35] is a normalized function, p_i represents position embedding, and s_t represents segment embedding. We set s_t to 1 in this work.

3.3.3. Image Embedding

The N independent image embedding obtained in Section 3.1 corresponds to the N tokens in the text modality. Firstly, we learn a randomly initialized weight matrix $W^m \in R^{2048 \times d}$ to project each 2048 dimensional image feature of the N image embeddings to the same d dimension as the text embedding, as shown in the following:

$$I_i = W^m f(image, i) \tag{11}$$

I_i represents the i-th output after the adaptive pooling layer. Then, we represent the visual features as $v = \{v_1, v_2, \ldots, v_N\} \in R^d$, where N represents N independent embedding of image features after the last adaptive pooling layer. Similarly, we add position embedding and segment embedding to get the final visual representation $\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_N\}$, and calculate the visual representation on the i-th position as follows:

$$\hat{v}_i = LayerNorm(v_i + p_i + s_v) \tag{12}$$

where s_v is set to 0.

3.4. Multimodal Encoder

After obtaining the sentence embedding vector and the visual embedding vector, we add two special tags [CLS] and [SEP] to construct the multimodal input sequence. [CLS] is used to learn the joint classification features, and [SEP] separates the embedding of different modality. The final input of the multimodal encoder is expressed as:

$$MF = \{[CLS], \hat{v_1}, \ldots, \hat{v_N}, [SEP], \hat{t}_1, \hat{t}_2, \ldots, \hat{t}_M\} \tag{13}$$

We then send the multimodal input $MF \in R^{d\times(M+N+2)}$ into the multimodal encoder based on transformer. The encoder contains a set of Bert layers for automatically modeling the rich interaction between textual and visual modality information. The architecture of Bert model is shown in Figure 3, where "Trm" stands for transformer [36], which is the infrastructure of Bert.

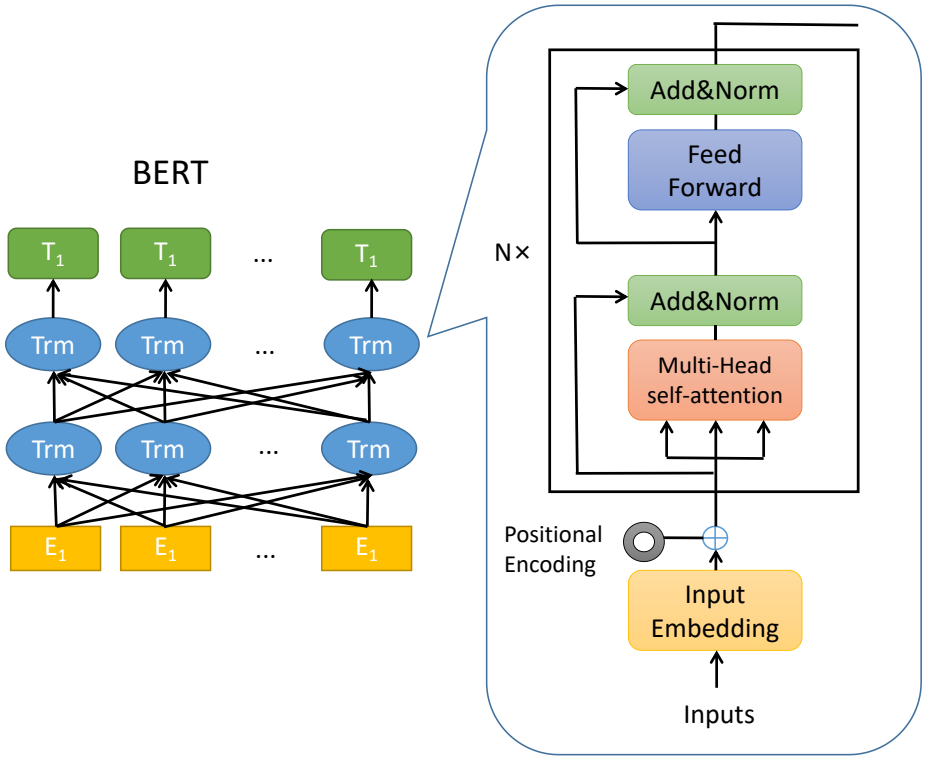


Figure 3. Architecture of the Bert model [36].

Firstly, the multimodal input MF through the multi-head self-attention layer pays attention to the information of different subspace to capture richer feature information. For the i-th head attention, the input $MF \in R^{d\times(M+N+2)}$ uses dot-product attention mechanism as follows:

$$Att_i(Q_i, K_i, V_i) = softmax(\frac{Q_i^T K_i}{\sqrt{d/m}})V_i \tag{14}$$

where m represents the number of heads, $Q_i = W_i^Q \cdot MF$, $K_i = W_i^K \cdot MF$, $V_i = W_i^V \cdot MF$ represent query, key, and value in i-th head attention, respectively. $W_i^Q \in R^{d\times d_Q}$, $W_i^K \in R^{d\times d_K}$, $W_i^V \in R^{d\times d_V}$ are three randomly initialized weight matrixes. In the BERT-base model, m is set to 12. $\sqrt{d/m}$ aims to turn the attention matrix into a standard normal distribution. Then, all attention heads are concatenated and multiplied by a weight matrix $W^o \in R^{d\times d}$ to obtain the output of multi-head self-attention, which is as shown as follows:

$$M(Q, K, V) = concat[Att_1(Q_1, K_1, V_1), \ldots, Att_m(Q_m, K_m, V_m)] \cdot W^o \tag{15}$$

Finally, residual connection and LayerNorm [35] operation are performed on the output of multi-head self-attention. The function of the LayerNorm operation aims to normalize the hidden layers in the neural network into standard normal distribution and accelerate convergence. Specific operations are as follows:

$$MF_{\text{attention}} = \text{Layernorm}\,(MF + M(Q, K, V)) \tag{16}$$

In addition, through the operation of two-layer linear mapping feed forward layer with Gelu [37] activation function and Equation (16), the output of an encoder in multimodal encoder is calculated as follows:

$$MF_{\text{out}} = \text{LayerNorm}[MF + \text{Gelu}(\text{Linear}(\text{Linear}(MF_{\text{attention}})))] \quad (17)$$

Then, MF_{out} is used as the input of the next multimodal encoder. The entire multimodal encoder stacks 12 of such encoder. Finally, the first token [CLS] of the last hidden layer is sent to the aesthetic prediction module to evaluate the image aesthetic distribution.

3.5. Aesthetic Prediction

After obtaining the multimodal feature vector through the above operations, we send it into a fully connected layer to output the aesthetic label distribution of the image. After that, the normalization operation will be carried out through the softmax layer. Similar with the work [13,14], this paper uses the Earth Mover's Distance (EMD) [19] loss function, which can calculate the minimum distance of the two sequential distributions, to optimize the network. The EMD loss is defined as follows:

$$\text{EMD}(p, \hat{p}) = \left(\frac{1}{N} \sum_{k=1}^{N} \left| CDF_p(k) - CDF_{\hat{p}}(k) \right|^r \right)^{1/r} \quad (18)$$

where p represents the real aesthetic score distribution of the image, and $\hat{p}$ represents the predicted aesthetic score distribution. $CDF(k)$ represents the cumulative distribution function. N represents the number of points. Similar to previous work [8,38], we choose $r = 2$ for its simplicity in optimization.

4. Experimental Results

In this section, we conduct a series of experiments on two benchmark datasets in the field of image aesthetics quality assessment to verify the effectiveness of our proposed method. First, the superiority of our method in multimodal fusion is explained by comparing with existing methods. Secondly, the effectiveness of the proposed method in dealing with the problem of missing modality data is proved by setting different levels of textual missing rates.

4.1. Experiment Setting

4.1.1. Dataset

AVA multimodal Dataset. AVA multimodal Dataset contains both images and text. The images are from the AVA Dataset [39], which is the largest and most widely used dataset in the field of image aesthetic quality assessment. It contains more than 250,000 photos, and each photo is scored by 200 users on average. The score is between 1–10. The higher the score, the higher the aesthetic quality of the image. The distribution of these scores is taken as the ground truth of our experiment. The text information comes from the AVA Comment dataset constructed by Zhang et al. [14], which contains users' comments on images. In addition, we use the method in [40] to further process the comment dataset and delete the over-long comments, over-short comments, and empty comments. After processing, we use 243,279 images for our experiment. The division method of training set and test set is consistent with [13,14]. In addition, we use 10% of the data in the training set as the validation set. Finally, the partition for the AVA database are 201,812 images for training, 22,431 for validation, and 19,036 images for test.

Photo.net multimodal Dataset. Photo.net multimodal Dataset is based on Photo.net Dataset, which is proposed by Ritendra Datta [15]. Each image in Photo.net dataset was rated by at least two members of the community. The scores are between 1.0 and 7.0. The text information comes from the photo.net comment dataset constructed by Zhang et al. [14]. They capture the user's comments on the image from the website. We also use the method in [40] to further process the comment dataset. Finally, the remaining

15,608 photos after processing are used in our experiment. The training set, validation set, and test set are 12,486, 1562, and 1560, respectively.

4.1.2. Evaluation Metrics

Existing methods usually formulate image aesthetic prediction task into three kinds: binary classification task [7,41] (i.e., distinguish images from high-quality to low-quality photos), regression task [1], and label distribution prediction task [8,14,19,29]. In order to compare with these methods fairly, we evaluate our proposed method on these three aesthetic quality tasks. The evaluation metrics corresponding to three tasks are as follows:

Aesthetic quality binary classification task. In image aesthetic quality classification tasks, we follow the experiment setup as [3]. If the average aesthetic score of the image is larger than 5, it will be regarded as a high-quality image, and if it is lower than 5, it will be regarded as a low-quality image. The definition accuracy is as follows:

$$\text{Accuracy} = \frac{TP + TN}{P + N} \tag{19}$$

Aesthetic score regression task. First, we predict the aesthetic score distribution of the image, and then calculate the mean score of the score distribution via score $= \sum_{i=1}^{N=10} s_i \times p_{s_i}$ to obtain the aesthetic regression scores for the regression task. The indices used to evaluate performance of the aesthetic quality regression task are the Spearman's Rank-ordered correlation coefficient ($SRCC$), Pearson linear correlation coefficient ($PLCC$), the Mean absolute error (MAE) and the root mean squared error ($RMSE$). $SRCC$ and $PLCC$ are the family of nonparametric correlation measures. $SRCC$ operates only on the rank of the data points, and measures the relative monotonicity between data-points, while $PLCC$ measures the linear association between the predictions and the subjective scores. Let y_i and s_i denote the prediction score and subjective score respectively, $SRCC$ and $PLCC$ are defined as follows:

$$SRCC = \frac{1 - 6\sum D^2}{n(n^2 - 1)} \tag{20}$$

where $D = (Y_i - S_i)$, Y_i and S_i are the rank order of y_i and s_i .

$$PLCC = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}} \tag{21}$$

where $\bar{y}$ and $\bar{s}$ represent the mean values of y_i and s_i. $RMSE$ and MAE measure the error between the real label and the predicted value. A better image aesthetic quality measurement has lower $RMSE$ and MAE values. As for a perfect match between the predicted scores and the subjective scores, $RMSE = MAE = 0$. $RMSE$ and MAE are defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - s_i)^2} \tag{22}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(|y_i - s_i|) \tag{23}$$

Aesthetic score distribution prediction task. Following the work of [13,14], we use the Earth mover's distance, which is defined defined in Equation (18) (r is set to 1) to judge the consistency between the predicted aesthetic distribution and the real aesthetic distribution.

4.1.3. Implementation Details

Our experiment is divided into two stages. In the first stage, we pretrain the missing modality generation model for missing modality reconstruction. In addition, in the second stage, we train our multimodal encoder module. We use the pytorch framework to build our model. Specifically, in the first stage, we pretrained the missing modality reconstruction

model on the DPC dataset and resize the image as 224 × 224. The output feature map is 14 × 14 × 512 after four pooling layers in VGG16. The learning rate of the image encoder is set to 0.0001. The dimension of word embedding m and LSTM n are both set to 512. For the LSTM as the decoder, the learning rate is set to 0.0004. The batch size is set to 32. In order to control the effect of gradient explosion, $grad_clip$ is set as 5. The length of the vocabulary k is set to 39,209.

In the second stage, the input image was resized to 224 × 224, and Resnet50 was initialized on the ImageNet dataset as the image encoder. We set the basic learning rate as 0.0001, $lr_patience$ as 5, and lr_factor as 0.1, that is, the learning rate will be reduced by 0.1 times when the performance has not improved after five epochs. The early stop strategy was adopted in the validation set. When the accuracy rate did not improve after 10 epochs, the training process was ended in advance. Due to the large amount of parameters in the experiment, we use the method of gradient accumulation to set *gradient_accumulation_steps* to 24, so that the batchsize can be set to a smaller 8. We use BertAdam with warmup rate of 0.1. The number of image feature embedding is set to 9. For the input text information, we specify a maximum length of 512, and fill it with zero if it is insufficient.

4.2. Ablation Studies

4.2.1. Effectiveness of Multimodal Features

In order to evaluate the effectiveness of multimodal features in the field of image aesthetic quality assessment, we compare the mutlimodal methods with single modality methods. The single modality methods include single column networks "Resnet50" and "VGG16" based on only image features, and also include single column networks "Bert" and "Text-CNN" based on only text features. The multimodal methods include double column networks "Resnet50+Bert" and "VGG16+Text-CNN" based on multimodal fusion features. The experimental results are shown in Table 1. We can find that: (1) the results obtained by using only text features are better than those obtained by using only image features. Specifically, the classification accuracy is improved by about 3% on average, and the improvement on SRCC is more significant, reaching 18.51% (BERT vs. VGG16). This may be because intuitive textual comments can better reflect people's aesthetic feelings about an image. (2) The use of multimodal features significantly improves the performance of each evaluation metrics. Specifically, the classification accuracy and SRCC were improved by 6.02% ("VGG16+Text-CNN" vs. "VGG16") and 16.77% ("Resnet50 + Bert" vs. "Resnet50"), respectively, compared with using only visual features, and 2.14% ("Resnet50 + Bert" vs. "Bert") and 4.4% ("VGG16 + Text-CNN" vs. "Text-CNN"), respectively, compared with using only text features. This shows the effectiveness of textual features and image features are complementary in the field of aesthetic assessment.

Table 1. Ablation study on the effect of multimodal features.

Model	Accuracy (%)↑	SRCC (Mean)↑	PLCC (Mean)↑	MAE↓	RMSE↓	EMD↓
Resnet50	79.25	0.6669	0.6736	0.4400	0.5639	0.048
VGG16	77.56	0.6267	0.6345	0.4659	0.5909	0.051
Bert	81.18	0.8118	0.8327	0.3328	0.4211	0.042
Text-CNN	81.85	0.7786	0.7851	0.3895	0.4887	0.046
Resnet50+Bert	83.32	0.8346	0.8495	0.3138	0.3984	0.039
VGG16+Text-CNN	83.58	0.8226	0.8365	0.3355	0.4032	0.038

4.2.2. Effectiveness of Missing Modality Reconstruction with Different Ratios

In order to evaluate the effectiveness of missing modality reconstruction, we designed a set of experiments. The missing modality in experiments randomly occurs in both training and testing phases. It is more general since we cannot be sure at which stage the textual modality missing occurs. The experimental results are shown in Table 2. In Table 2, "empty" indicates that the textual modality of the samples is missing. We only input the available visual modality information for training and testing. "Reconstruct" is used to represent that we reconstruct the missing textual modality according to the available visual modality information. Finally, the reconstructed textual modality is combined with the available visual modality for training and testing. From the table, we can observe that the performance is improved to varying degrees under different textual modality missing rate after missing modality reconstruction. For example, with the 20% missing rate, the classification accuracy of improves from 82.58% to 83.15%, the SRCC improves from 0.8050 to 0.8204, and the PLCC improves from 0.8186 to 0.8398. The experimental results show that the superior performance of the missing modality reconstruction model is essential to deal with the missing modality problem.

We also performed some qualitative analysis to visually see the effectiveness of the generated text modality. Some examples are shown in Figure 4. We use some full-modality samples of the AVA dataset for testing, and compare the actual text representation with the generated text. From the figure, we can easily find that our model can reconstruct most key words.

Table 2. Ablation study on the effect of textual modality reconstruction.

Missing_Rate	Text_State	Accuracy (%)↑	SRCC (Mean)↑	PLCC (Mean)↑	MAE↓	RMSE↓	EMD↓
10%	empty	83.44	0.8327	0.8449	0.3149	0.4078	0.036
	reconstruct	83.73	0.8498	0.8628	0.3014	0.3818	0.036
20%	empty	82.58	0.8050	0.8186	0.3341	0.4343	0.038
	reconstruct	83.15	0.8204	0.8398	0.3243	0.4232	0.037
30%	empty	81.91	0.7798	0.7941	0.3532	0.4621	0.043
	reconstruct	82.38	0.7998	0.8121	0.3342	0.4342	0.039



Figure 4. Some examples of AVA datasets. "groundtruth" represents the tag value, "binary classification" represents the result of binary classification, "regression score" represents the regression score of the image (keep one decimal place), "comment" represents the real comment, "generate" represents the text comment generated by the missing modality generation network.

4.2.3. The Effectiveness of Multimodal Encoder Based on Transformer

We compare the performance of transformer-based multimodal encoder with other multimodal fusion methods, i.e., feature concatenation and MFB. These two methods are often used in existing multimodal image aesthetic prediction task. In order to make a fair comparison, the experimental settings on these two baseline methods are consistent with the experimental settings of the method proposed in this paper.

Feature concatenation. We use Resnet50 to remove the last two layers to extract image features, and use Bert to extract text features. Finally, the multimodal features are concatenated and sent into a fully connected layer to predict the aesthetic distribution. In the specific experiment, we set the length of the image feature vector to 2048 and the length of the text feature vector to 768. The multimodal feature vector with the length of 2816 is sent to the full connected layer with the output dimension of 10, and a softmax activation function is added to predict the aesthetic distribution.

MFB. Multimodal Factorized Bilinear pooling (MFB) can encode the complex interactions between features and thus was frequently used in existing multimodal fusion methods [14,42]. In order to make a fair comparison, we also use ResNet50 to extract visual features and use Bert to extract textual features. Then, these two features are sent into MFB for further processing. In the specific experiment, we wrap the 2048-dimensional image feature vector input to 768 dimensions through a fully connected layer. Then, the 768 dimension textual feature and image features are sent into the MFB for fusion.

The experimental results are shown in Table 3. We can clearly observe that our method is superior to the other two multimodal fusion methods in all evaluation metrics. It proves that the transformer-based multimodal encoder can effectively fuse multimodal features, thereby improving the performance of the image aesthetics quality assessment task.

Table 3. Ablation study on the effect of the Multimodal encoder based on transformer.

Model	Accuracy (%)↑	SRCC (Mean)↑	PLCC (Mean)↑	MAE↓	RMSE↓	EMD↓
Feature concatenation	83.32	0.8346	0.8495	0.3138	0.3984	0.036
MFB	83.64	0.8340	0.8450	0.3129	0.4044	0.036
Ours	**84.05**	**0.8511**	**0.8656**	**0.2963**	**0.3788**	**0.034**

We also found that the proposed multimodal fusion module has the advantages in dealing with different ratios of modality missing. The results on AVA dataset are shown in Table 4. We set three different modality missing rate, which are 10%, 20%, and 30%, respectively. As can be seen, our approach outperforms all baselines among all different ratios of textual modality missing, which showcases the efficiency of our method in the missing textual modality problem. More specifically, under different textual modality missing rates, our method is higher than concatbert and MFB by about 1.5% in accuracy, higher than concatbert and MFB by about 5% in SRCC and PLCC evaluation metric, and lower than concatbert and MFB by about 3% in MAE and RMSE evaluation metric.

Table 4. Comparison to different multimodal fusion methods under three textual missing modality ratios (10%, 20%, and 30%) on the AVA dataset.

Missing_Rate	Method	Accuracy (%)↑	SRCC (Mean)↑	PLCC (Mean)↑	MAE↓	RMSE↓	EMD↓
10%	Feature concatenation	82.47	0.8089	0.8251	0.3314	0.4278	0.037
	MFB	82.49	0.8159	0.8262	0.3303	0.4290	0.036
	ours	**83.73**	**0.8498**	**0.8628**	**0.3014**	**0.3818**	**0.034**
20%	Feature concatenation	81.52	0.7909	0.7979	0.3507	0.4554	0.038
	MFB	82.09	0.7767	0.7978	0.3478	0.4588	0.038
	ours	**83.15**	**0.8204**	**0.8398**	**0.3243**	**0.4232**	**0.037**
30%	Feature concatenation	80.43	0.7802	0.7859	0.3689	0.4728	0.047
	MFB	80.98	0.7594	0.7759	0.3682	0.4769	0.045
	ours	**82.38**	**0.7998**	**0.8121**	**0.3342**	**0.4342**	**0.039**

4.2.4. Extension to Different Image Encoder

Our proposed model does not depend on a specific image encoder, it can accept any dense sequence as input. In order to compare the impact of different image encoders on the performance of the proposed model, we choose Resnet50 [27], VGG16 [43], and Densenet161 [44]—these three representative convolutional neural networks are our image encoder. Resnet50 [27] introduces residual connections, improves the flow of information, and solves the problem of vanishing gradient and degradation caused by too deep a network. VGG16 [43] has a relatively simple structure. Densenet161 [44] has a narrower network structure and fewer parameters, which can enhance the transmission of features. The three networks are pretrained on the ImageNet dataset, and then fine-tuned on the AVA dataset. The experimental results are shown in Table 5. We can find that the performance of the proposed model on the three CNN is similar, but the effect is better than that of the baseline.

Table 5. Performance of different image encoder architecture.

Model	Accuracy (%)↑	SRCC (Mean)↑	PLCC (Mean)↑	MAE↓	RMSE↓	EMD↓
Resnet50	79.25	0.6669	0.6736	0.4400	0.5639	0.048
VGG16	77.56	0.6267	0.6345	0.4659	0.5909	0.051
Densenet	78.94	0.6466	0.6481	0.4550	0.5814	0.049
ours(Resnet50)	**84.05**	**0.8511**	**0.8656**	**0.2963**	**0.3788**	**0.034**
ours(VGG16)	**83.87**	**0.8528**	**0.8677**	**0.2904**	**0.3752**	**0.034**
ours(Densenet161)	**84.32**	**0.8528**	**0.8683**	**0.2928**	**0.3759**	**0.033**

4.2.5. Effects of the Various Image/Text Embedding Lengths

The length of image embeddings and text embeddings are important factors affecting the image aesthetic quality prediction task. Therefore, we compared the performance of our proposed model under different lengths of image embeddings and text embeddings. The experimental results are shown in Table 6, from which we can clearly see that the performance was best performed when image embedding length is 2048, and text embedding length is 768. Thus, we set the length of image embedding as 2048 and the length of text embedding as 768.

Table 6. Performance comparison with various image/text embedding lengths.

Image Embedding Length	Text Embedding Length	Accuracy (%)↑	SRCC (Mean)↑	PLCC (Mean)↑	MAE↓	RMSE↓	EMD↓
1024	512	83.81	0.8512	0.8533	0.3013	0.3900	0.035
1024	768	83.78	0.8499	0.8507	0.3011	0.3894	0.036
2048	512	83.91	0.8505	0.8599	0.2986	0.3812	0.035
2048	768	**84.05**	**0.8511**	**0.8656**	**0.2963**	**0.3788**	**0.034**

4.3. Comparison to State-of-the-Art Methods on the AVA Dataset

We compared the proposed model with other related work in the field of image aesthetics quality evaluation, and we chose the following nine typical methods. Among them, the first six methods all use a single image modality for training. While the rest methods, such as Joint DBM [12], MultiGap [22], and SAGAN [38] are multimodal methods. However, none of them consider the missing modality problem.

RAPID [3]: RAPID [3] is the first work that tried to apply convolutional neural networks to this field. It consists of a double-column deep convolutional neural network to learn the global and local features of the image respectively. In this work, image aesthetic was formulated as a binary classification task. Thus, we only need the accuracy metric to evaluate the performance.

MTCNN [41]: It is an end-to-end multi-task deep learning framework that adds semantic information to perform a binary image aesthetic classification task.

Full model [45]: In [45], Xu et al. proposed a context-aware attention-based image aesthetic score prediction method. The context-aware attention module is in two dimensions: hierarchical and spatial. The hierarchical context aims to encode the multi-level aesthetic details while the spatial context encodes the long-range perception of images.

NIMA [19]: In [19], an aesthetic distribution prediction task is proposed, which can better reflect human subjective performance of images. In addition, EMD loss function is introduced into this field and greatly improved the performance.

ARDP [46]: An object-level attention based prediction framework of aesthetic grade distribution is proposed. The framework dynamically learns the features extracted from the object level region defined by the general object detector.

GPF-CNN [8]: The proposed architecture can extract fine detail features and adaptive fuse global and local features according to the input feature map.

Joint DBM [12]: For the first time, it considered to use both image information and text information to improve the performance of the image aesthetic quality assessment. Firstly, different modality features are extracted with different network architectures, and then multimodal joint representation is learned by DBM.

MultiGap [22]: It uses inception to extract image features and RNN to extract text features. Finally, the two features are directly concatenated to complete the task of aesthetic binary classification.

SAGAN [38]: SAGAN [38] makes full use of the intrinsic relationship between aesthetic attributes and aesthetics through the process of semi-supervised attribute learning and adversarial training.

The experimental results are shown in Table 7. It can be found that the proposed model is superior to other methods on all three tasks of image aesthetic quality assessment. Compared with the earlier unimodal methods, the performance of the proposed model is 10.12% and 5.86% higher accuracy than RAPID [3] and MTCNN [41], respectively. For Full model [45], NIMA [19], ARDP [46], and GPF-CNN [8], which use EMD loss to predict the aesthetic score distributions of the image, our method is higher than them in various evaluation metrics. This demonstrates the superiority of the proposed method. The recently proposed method SAGAN exploits high-level attributes to improve the aesthetic prediction

performance, and thus outperforms NIMA [19], ARDP [46], and GPF-CNN [8], achieving 83.72% in the aesthetic classification task. However, our model achieves the strongest results at an accuracy rate of 84.32%. Joint DBM [12] and MultiGap [22] are closely related to our method since they both use the textual and visual information to predict aesthetics. MultiGap outperforms a multimodal DBM model, achieving 82.27% accuracy. In contrast, our method achieves 84.32% in classification accuracy and 0.8528 in SRCC. Even the text modality missing rate is 20%, and the accuracy of our method is still higher than these two models, which further show the superiority of our method.

Table 7. Comparison with state-of-the-art methods on the AVA dataset.

Method	Accuracy (%)↑	SRCC (Mean)↑	PLCC (Mean)↑	MAE↓	RMSE↓	EMD↓
RAPID	74.2	-	-	-	-	-
MTCNN	78.46	-	-	-	-	-
Full Model	80.9	0.7240	0.7250	-	-	0.052
NIMA	80.6	0.592	0.610	-	-	0.052
ARDP	81.67	0.7510	0.7530	-	-	0.052
GPF-CNN	80.70	0.6762	0.6868	0.4144	0.5347	0.046
Joint DBM	78.88	-	-	-	-	-
MultiGap	82.27	-	-	-	-	-
SAGAN	83.72	0.774	0.788	-	-	0.052
ours	**84.32**	**0.8528**	**0.8683**	**0.2928**	**0.3759**	**0.033**
ours (10%)	83.73	0.8498	0.8628	0.3014	0.3818	0.034
ours (20%)	83.15	0.8204	0.8398	0.3243	0.4232	0.037
ours (30%)	82.38	0.7998	0.8121	0.3342	0.4342	0.039

4.4. Comparison to State-of-the-Art Methods on the Photo.Net Dataset

We also compare our proposed method with several existing models on the Photo.net dataset. The experimental results are shown in Table 8. *GLST_SVM* and *FV_SIFT_SVM* [47] use handcrafted features to predict the image aesthetics. MTCNN [41] and GPF-CNN [8] are single-mode methods which rely on visual information to make aesthetic decisions. MRACNN [14] is closely related to our method, which both use the textual and visual information to jointly predict the image aesthetic distribution. From the table, we can see that our proposed model keeps the best and achieves 79.18% for aesthetic classification accuracy and 0.6553 on the SRCC metric.

Table 8. Comparison with state-of-the-art methods on the Photo.net Dataset.

Method	Accuracy (%)↑	SRCC (Mean)↑	PLCC (Mean)↑	MAE↓	RMSE↓	EMD↓
GIST_SVM	59.90	-	-	-	-	-
FV_SIFT_SVM	60.8	-	-	-	-	-
MTCNN	65.2	-	-	-	-	-
GPF-CNN	75.6	0.5217	0.5464	0.4242	0.5211	0.070
MRACNN	78.91	0.5709	0.5902	0.3636	0.4589	**0.0622**
ours	**79.18**	**0.6553**	**0.6670**	**0.3263**	**0.4181**	**0.054**
ours (10%)	78.59	0.6469	0.6626	0.3348	0.4267	0.058
ours (20%)	76.81	0.6142	0.6350	0.3414	0.4309	0.064
ours (30%)	75.93	0.5688	0.5909	0.3576	0.4556	0.069

5. Conclusions

In this paper, we propose a novel multimodal image aesthetic quality assessment method, which not only solves the problem of missing textual modality, but also breaks through the limitation of feature fusion only at a single level in the previous methods. This method combines the most advanced methods in the field of computer vision and natural language processing. The pretrained missing modality reconstruction model reconstructs the missing textual modality according to the available visual modality, thereby forming a new multimodal representation. The proposed multimodal encoder can make multimodal information interact at different levels for more effective and fine-grained fusion. Experimental results on AVA and Photo.net datasets show that our method not only improves the performance in full-modality conditions, but also can effectively solve the problem of missing textual modality. In the future, we will explore more effective solutions for severely missing modality problems.

Author Contributions: Conceptualization, X.Z. and Q.S.; methodology, X.Z.; software, X.Z. and Q.S.; validation, X.Z. and G.L.; formal analysis, G.L.; resources, G.L.; writing—original draft preparation, X.Z. and Q.S.; writing—review and editing, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant No. 62001385, in part by the Key RD Program of Shaanxi under Grant 2021ZDLGY15-03, and in part by the Project funded by China Postdoctoral Science Foundation (Grant No. 2021MD703883).

Institutional Review Board Statement: Not applicable .

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in [15,39].

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Kong, S.; Shen, X.; Lin, Z.; Mech, R.; Fowlkes, C. Photo aesthetics ranking network with attributes and content adaptation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2016; pp. 662–679.
2. Sun, W.T.; Chao, T.H.; Kuo, Y.H.; Hsu, W.H. Photo filter recommendation by category-aware aesthetic learning. *IEEE Trans. Multimed.* **2017**, *19*, 1870–1880. [CrossRef]
3. Lu, X.; Lin, Z.; Jin, H.; Yang, J.; Wang, J.Z. Rapid: Rating pictorial aesthetics using deep learning. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 457–466.
4. Guo, L.; Xiong, Y.; Huang, Q.; Li, X. Image esthetic assessment using both hand-crafting and semantic features. *Neurocomputing* **2014**, *143*, 14–26. [CrossRef]
5. Luo, W.; Wang, X.; Tang, X. Content-based photo quality assessment. In Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2206–2213.
6. Nishiyama, M.; Okabe, T.; Sato, I.; Sato, Y. Aesthetic quality classification of photographs based on color harmony. In Proceedings of the CVPR, Colorado Springs, CO, USA, 20–25 June 2011; pp. 33–40.
7. Lu, X.; Lin, Z.; Jin, H.; Yang, J.; Wang, J.Z. Rating image aesthetics using deep learning. *IEEE Trans. Multimed.* **2015**, *17*, 2021–2034. [CrossRef]
8. Zhang, X.; Gao, X.; Lu, W.; He, L. A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction. *IEEE Trans. Multimed.* **2019**, *21*, 2815–2826. [CrossRef]
9. Jin, B.; Segovia, M.V.O.; Süsstrunk, S. Image aesthetic predictors based on weighted CNNs. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 2291–2295.
10. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
11. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 818–833.
12. Zhou, Y.; Lu, X.; Zhang, J.; Wang, J.Z. Joint image and text representation for aesthetics analysis. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 262–266.
13. Zhang, X.; Gao, X.; He, L.; Lu, W. MSCAN: Multimodal Self-and-Collaborative Attention Network for image aesthetic prediction tasks. *Neurocomputing* **2021**, *430*, 14–23. [CrossRef]

14. 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. *IEEE Trans. Multimed.* **2020**, *23*, 611–623. [CrossRef]
15. 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, Graz, Austria, 7–13 May 2006; pp. 288–301.
16. Aydın, T.O.; Smolic, A.; Gross, M. Automated aesthetic analysis of photographic images. *IEEE Trans. Vis. Comput. Graph.* **2014**, *21*, 31–42. [CrossRef]
17. Hulusic, V.; Valenzise, G.; Provenzi, E.; Debattista, K.; Dufaux, F. Perceived dynamic range of HDR images. In Proceedings of the 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, Portugal, 6–8 June 2016; pp. 1–6.
18. Kao, Y.; Wang, C.; Huang, K. Visual aesthetic quality assessment with a regression model. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 1583–1587.
19. Talebi, H.; Milanfar, P. NIMA: Neural image assessment. *IEEE Trans. Image Process.* **2018**, *27*, 3998–4011. [CrossRef]
20. Wang, L.; Wang, X.; Yamasaki, T.; Aizawa, K. Aspect-ratio-preserving multi-patch image aesthetics score prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
21. Zheng, X.; Wang, B.; Du, X.; Lu, X. Mutual attention inception network for remote sensing visual question answering. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *60*, 1–14. [CrossRef]
22. Hii, Y.L.; See, J.; Kairanbay, M.; Wong, L.K. Multigap: Multi-pooled inception network with text augmentation for aesthetic prediction of photographs. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 1722–1726.
23. Miao, H.; Zhang, Y.; Wang, D.; Feng, S. Multi-Output Learning Based on Multimodal GCN and Co-Attention for Image Aesthetics and Emotion Analysis. *Mathematics* **2021**, *9*, 1437. [CrossRef]
24. Tran, L.; Liu, X.; Zhou, J.; Jin, R. Missing modalities imputation via cascaded residual autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1405–1414.
25. Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; Peng, X. Smil: Multimodal learning with severely missing modality. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 11–15 October 2021; Volume 35, pp. 2302–2310.
26. Zhang, C.; Fu, H.; Zhou, J.T.; Hu, Q. CPM-Nets: Cross partial multi-view networks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
28. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
29. Jin, X.; Wu, L.; Zhao, G.; Li, X.; Zhang, X.; Ge, S.; Zou, D.; Zhou, B.; Zhou, X. Aesthetic attributes assessment of images. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 311–319.
30. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
31. Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent neural network regularization. *arXiv* **2014**, arXiv:1409.2329.
32. Pascanu, R.; Gulcehre, C.; Cho, K.; Bengio, Y. How to construct deep recurrent neural networks. *arXiv* **2013**, arXiv:1312.6026.
33. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
34. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.
35. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
37. Hendrycks, D.; Gimpel, K. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *arXiv* **2016**, arXiv:1606.08415.
38. Shu, Y.; Li, Q.; Liu, L.; Xu, G. Semi-supervised Adversarial Learning for Attribute-Aware Photo Aesthetic Assessment. *IEEE Trans. Multimed.* **2021**. [CrossRef]
39. Murray, N.; Marchesotti, L.; Perronnin, F. AVA: A large-scale database for aesthetic visual analysis. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2408–2415.
40. Ghosal, K.; Rana, A.; Smolic, A. Aesthetic image captioning from weakly-labelled photographs. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.
41. Kao, Y.; He, R.; Huang, K. Deep aesthetic quality assessment with semantic information. *IEEE Trans. Image Process.* **2017**, *26*, 1482–1495. [CrossRef] [PubMed]
42. Yu, Z.; Yu, J.; Xiang, C.; Fan, J.; Tao, D. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 5947–5959. [CrossRef] [PubMed]
43. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
44. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.

45. Xu, M.; Zhong, J.X.; Ren, Y.; Liu, S.; Li, G. Context-aware attention network for predicting image aesthetic subjectivity. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 798–806.
46. Hou, J.; Yang, S.; Lin, W. Object-level attention for aesthetic rating distribution prediction. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 816–824.
47. Marchesotti, L.; Perronnin, F.; Larlus, D.; Csurka, G. Assessing the aesthetic quality of photographs using generic image descriptors. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1784–1791.