This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2985526, IEEE Transactions on Multimedia

Beyond Vision: A Multimodal Recurrent Attention Convolutional Neural Network for Unified Image Aesthetic Prediction Tasks

Xiaodan Zhang, Xinbo Gao, *Senior Member, IEEE,* Wen Lu, Lihuo He, and Jie Li

***Abstract*—Over the past few years, image aesthetic prediction has attracted increasing attention because of its wide applications, such as image retrieval, photo album management and aesthetic-driven image enhancement. However, previous studies in this area only achieve limited success because 1) they primarily depend on visual features and ignore textual information. 2) they tend to focus equally on to each part of images and ignore the selective attention mechanism. This paper overcomes these limitations by proposing a novel multimodal recurrent attention convolutional neural network (MRACNN). More specifically, the MRACNN consists of two streams: the vision stream and the language stream. The former employs the recurrent attention network to tune out irrelevant information and focuses on some key regions to extract visual features. The latter utilizes the Text-CNN to capture the high-level semantics of user comments. Finally, a multimodal factorized bilinear (MFB) pooling approach is used to achieve effective fusion of textual and visual features. Extensive experiments demonstrate that the proposed MRACNN significantly outperforms state-of-the-art methods for unified aesthetic prediction tasks: (i) aesthetic quality classification; (ii) aesthetic score regression; and (iii) aesthetic score distribution prediction.**

***Index Terms*—Image quality assessment, visual aesthetic quality assessment, long short-term memory (LSTM), deep learning**

I. Introduction

With the ever-expanding volume of digital photographs on the Internet, automatic image aesthetic prediction has become increasingly in many applications, such as image recommendation, consumer photography, and image enhancement [1]–[3]. Thus, it has gained a great interest from the research community. Compared with traditional image quality assessment methods [4], [5], which focus on low-level perceptual degradation, image aesthetic assessment is much more challenging since it needs to understand the high-level abstraction of the image and consider human subjectivity in the process of aesthetic perception.

Manuscript received August 2, 2019; revised January 27, 2020; accepted March 30, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61432014, 61772402, U1605252 and 61671339, in part by the National Key Research and Development Program of China under Grant 2016QY01W0200, and in part by National High-Level Talents Special Support Program of China under Grant CS31117200001.(*Corresponding author: Xinbo Gao.*)

X. Zhang is with School of Information Science and Technology, Northwest University, Xi'an, Shaanxi, China. X. Gao, W. Lu, L. He and Jie Li are with the Video and Image Processing System Laboratory, School of Electronic Engineering, Xidian University, Xi'an 710071, China (email: xdanzhang.opt@gmail.com; xbgao@mail.xidian.edu.cn; luwen@mail.xidian.edu.cn; lihuo.he@gmail.com; leejie@mail.xidian.edu.cn).

(a) Nice lighting and composition. Good capture of detail.

(b) This is fabulous! Outstanding on so many levels: HDR, composition, focus, color! Beautiful art!

(c) The overshapening around the tree is distracting.

(d) The bits of yellow in the left part of the frame could have been cropped out. Also, it is a bit dark on the flowers.

Fig. 1. Example images from AVA datasets with their associated comments. Top row: high-quality images. Bottom row: low-quality images..

Earlier attempts have followed the photographic rules by defining features that capture low-level cues such as colour, texture and contrast [6]–[8] or high-level features such as image composition, image saliency and scene contexts [9]. However, these handcrafted features are heuristic, and the representation ability is limited. Later, other approaches leveraged the generic image features (*e.g.*, bag-of-visual words [10] and Fisher vector [11]) to the aesthetic prediction. Although generic features show promising performance compared with rule-based methods, they may not be optimal for photoaesthetics because they are designed to encode the general characteristics of natural images, not specifically for aesthetic prediction [12]. Recently, deep learning methods have shown great success in various computer vision tasks [13], [14]. An increasing number of researchers have attempted to apply deep

learning methods to image aesthetic assessment [15]–[17]. However, most existing methods tend to generate a global image representation by focussing equally on each region despite the irrelevant areas. Furthermore, they primarily focus on extracting vision features from images, which is quite insufficient for image aesthetic prediction.

On many image sharing websites, *e.g.*, Flickr, DPChallenge, users not only give scores but also explain why they gave such a score. Fig 1 illustrates some examples. These user comments encode high-level semantic information and are quite important in aesthetic prediction. Comments such as "nice lighting and composition", "outstanding focus", or "beautiful art" indicate that the images may be a professional photograph. Similarly, comments such as "distracting", or "a bit dark" may infer that the images may be snapshots.

In this paper, we propose a multimodal recurrent attention network for image aesthetic prediction. The proposed network not only aims to extract more effective visual features but also leverages the semantic power of user comments to make aesthetic predictions more accurate. It consists of three components: a discriminative visual feature extractor, a textual feature extractor, and a multimodal feature fusion module. For the visual feature extractor, directly using the global features extracted from the whole image may introduce noisy information [12]. Randomly cropping methods such as [18], [19] may undermine the integrity of semantic information and change image composition. However, in our visual attention mechanism, we do not process the entire image at once. Instead, we iteratively select some important regions and combine them to generate an internal representation. Inspired by this, we use a recurrent attention network to generate several attention representations based on the high-level features extracted from CNN. Finally, these attentional representations are concatenated to generate the visual features. For the textual feature extraction module, the user comments are first cleaned and then mapped into a rich high-dimensional space using word embeddings. The embedded vectors are then fed to a Text-CNN [20] network to capture the semantics. For the multimodal feature fusion module, most methods adopt linear fusion approaches, such as concatenation or elementwise summarization or production [21], [22]. Such linear fusion methods fail to encode the complex relationship between the multimodal features. In contrast, we adopt multimodal factorized bilinear pooling methods to model the correlations between different multimodal features [23], which can achieve superior performance.

Overall, the main contributions are as follows.

- A novel end-to-end multimodal network is proposed to jointly learn the visual features and textual features for image aesthetic prediction. The complex interactions between the multimodal features are captured via the multimodal bilinear pooling method. Experimental results demonstrate that integrating visual and textual features can significantly boost performance.
- Inspired by the human attention mechanism, a recurrent attention neural network is used to extract visual features. It can learn to iteratively focus on important regions and tune out irrelevant information. To the best of our knowledge, this is the first successful attempt to elucidate the recurrent attention mechanism for image aesthetic assessment.
- We collect the AVA comment dataset and the photo.net comment dataset. These datasets can advance the research on multimodal modelling in image aesthetics.
- Comprehensive experiments for unified aesthetic prediction tasks, e.g., aesthetic classification, aesthetic regression and aesthetic label distribution, are conducted. For all these tasks, the proposed model achieves superior performance over the state-of-the-art approaches.

The remainder of this paper is organized as follows. Section II briefly summarizes the related work and Section III introduces the architecture of the MRACNN model. Section IV quantitatively evaluates the effectiveness of the proposed model and compares it with state-of-the-art methods. Finally, conclusions and ideas for future work are wrapped up in Section V.

II. Related Work

In this section, we briefly review the relevant work from the following three research topics: 1) computational image aesthetics prediction, 2) cross-modal analysis. 3) attention mechanism

A. Computational Image Aesthetics Prediction

Existing image aesthetic assessment research can be roughly outlined by the following two important components: extraction of more advanced features and utilization of more sophisticated learning algorithms.

1) *Features Representations.* Many different communities have tried to address the problem of image aesthetic prediction such as [7], [8], [24]–[28]. Early pioneering methods are mainly focused on handcrafted features. Among these features, exposure of light, colorfulness, saturation and hue [7], [8] are global features that can be used for all the image categories. Dark channel feature, clarity contrast feature, and composition geometry feature are local features which need to be designed according to the variety of photo content [6], [7]. Generic features such as SIFT, HOG and Fisher Vector [11] can also be used to predict photo aesthetics. However, these handcrafted features are heuristic and the representation ability is limited.

Recently, some researchers have tried to apply the deep learning networks to image aesthetic quality assessment. Kao *et al.* [26] use the image tags as extra supervision and propose a deep multi-task network to predict the image aesthetics. Kucer *et al.* [27] incorporate the hand-designed features with the deep learned features to further boost the performance. Different from the above mentioned methods [18], [19], [29], [30] focus on encoding the global composition and local fine-details. Lu *et al.* [19] propose the *RAPID* to support the heterogeneous inputs, *i.e.,* the global views and the local views. The global view is represented with a warped image, and the local view is represented with a randomly cropped patch. Similarly, Zhang *et al.* [30] also propose a double-subnet neural network, but they extract the fine-grained information with the guide of top-down neural attention. In order to capture

more high-resolution fine-grained details, [18] and [29] further propose a deep multi-patch aggregation network (*DMA-Net*). *DMA-Net* represents the input images with a bag of random cropped patches, and use two network layers (statistics and sorting) to aggregate these multiple patches. The *DMA-Net* architecture significantly outperforms the results of baselines. However, the random cropping strategy in *DMA-Net* may undermine the integrity of semantic information. Our model is different from *DMA-Net* fundamentally. The difference can be summarized in two aspects. 1) Based on the random cropping strategy, there is no guarantee that the *DMA-Net* well locates the important regions to extract visual features. Comparatively, our approach uses the recurrent attention mechanism to iteratively extract visual features on the key regions for further processing. 2) The *DMA-Net* simply captures the visual features of a photograph. While our model jointly learns the visual and textual features together, resulting in the significant improvements in unified aesthetic prediction tasks, *i.e.*, image aesthetic classification, image aesthetic regression, and image aesthetic label distribution.

2) *Learning Algorithms.* Image aesthetic assessment is quite a complex and subjective task. Early attempts in this area often cast this problem as a classification task, such as [6], [7], [18], [26], [27], [29]. They classify the images into two categories: *i.e.* high quality and low quality images, and design leaning algorithm to minimize classification error. Other methods formulate image aesthetic prediction as a regression task, such as [16], [17]. But directly regressing the mean score is a difficult problem. The images in the aesthetic dataset are often rated by people from different cultural backgrounds. Consequently, the aesthetic annotations usually come in the form of a distribution of scores (e.g. from $1-10$ in AVA dataset and from $1-7$ in Photo.net dataset.). However, the traditional classification [6], [7], [18], [26]–[29] and regression methods [16], [17], [31] transform the rating distributions into a scalar value and remove the distribution information. The removed information indicates the consensus or diversity of human opinions, and is quite important to show the subjectivity of aesthetic prediction. Some researchers have noted this limitation and focus on directly predicting the label distribution of the scores [30], [32]–[34]. In [32], Murray *et al.* propose to use the Huber loss to predict the aesthetic score distribution. But they predict each discrete probability independently and ignore their relationships. Comparatively, Talebi *et al.* [33] and Zhang *et al.* [30] treat the score distribution as ordered classes and used squared EMD (Earth Mover's Distance) loss to predict the score distributions.

B. Cross-Modal Analysis

With the rapid growth of multimedia information, multiple kinds of modalities that describe the same content can be easily obtained, such as audio, images, and text. These modalities are closely related and can provide complementary information for each other. Multi-modal approaches have outperformed single-modal in many tasks. For example, in [35] and [22], the authors jointly combine the vision and language to improve the performance of fine-grained image classification. Hu *et al.* [21] adaptively fuse the image features and textual descriptions to explore the structure of emotions in visual sentiment prediction.

Although multimodal methods have achieved good performance in many tasks, joint multimodal learning has been rarely addressed in image aesthetic prediction, with a few exceptions [23], [36]. Zhou *et al.* [36] are the first to incorporate additional textual features to represent image aesthetics. They build a multimodal Deep Boltzmann Machines (DBM) to perform the joint representation learning. Furthermore, Wang *et al.* [37] use textual reviews to guide the personalized image aesthetic prediction. The main challenge of multimodal approaches lies in how to optimally combine the multimodal information. In this paper, we adopt the MFB pooling approach [23] to model the high-order interactions between the features. Experimental results demonstrate it can achieve significant improvement on the aesthetic prediction tasks.

C. Attention Mechanism

Visual attention allows us to efficiently deal with complex scenes by selecting relevant information and by filtering out irrelevant information. Considering its widely applications, a large quantity of visual attention models has been proposed in the literature, such as [38], [39]. The potential benefit of integrating visual attention mechanism into image aesthetic models has recently been recognized by a number of research groups. For example, Sun et al. [40] estimate visual attention distribution to describe an image, and then predict the aesthetic score of an image based on the rate of focused attention region in the saliency map. Sheng et al. [41] use the attention mechanism to select patches during the training process. Although these methods achieve impressive performance, they just use the saliency to sampling and neglect the human perception process. When human perceive the scenes, they sequentially attend to different parts of the visual space to acquire useful information, and combine information from different fixations to build up an internal representation of the scene [39]. In this paper, we fill the gap by adopting the recurrent attention network to emulate this process. The experimental results demonstrate it can achieve better performance than traditional methods.

III. Multimodal Recurrent Attention Convolutional Neural Network

Our model is based on a very simple intuition: user comments provide the textual aesthetic attributes, and such information is complementary with visual features in aesthetic decisions. Therefore, we propose a two-stream model combining vision and language. In the vision stream, images are fed into the CNN architecture to extract high-level features. Recurrent attention network is adopted for iteratively focusing on key regions to generate the attentional representation. As for the language stream, user comments are tokenized and fed into the Text-CNN to extract textual semantic information. Finally, the pairwise interactions between visual and textual features are encoded via a MFB pooling module. The overall architecture of the model is shown in Fig. 2. In order to fairly

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2985526, IEEE Transactions on Multimedia

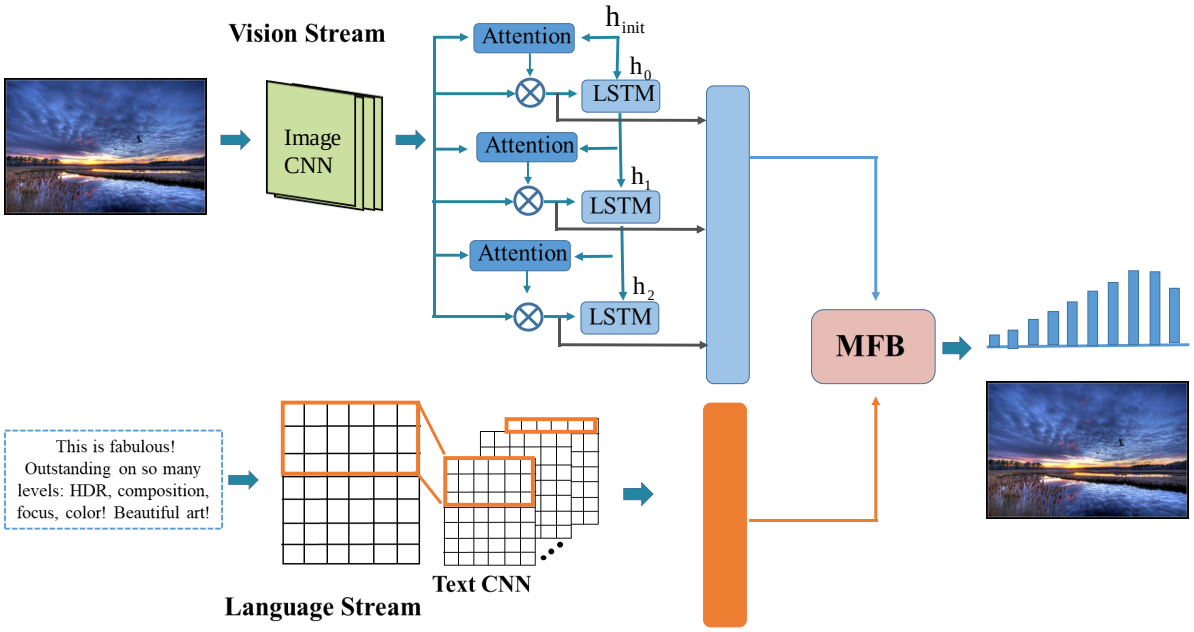


Fig. 2. Overall architecture of the MRACNN. The vision stream employs the recurrent attention network to learn visual features on important regions. The language stream uses the Text-CNN to extract semantic information from user comments. MFB is used to fuse the multimodal features. More detailed illustrations of MFB module can be found in Fig. 5.

compare with previous works such as [30], [33], we adopt the EMD loss which is defined as follows:

$$EMD(q,\hat{q}) = (\frac{1}{N}\sum_{k=1}^{N}|CDF_q(k) - CDF_{\hat{q}}(k)|^r)^{\frac{1}{r}}, \quad (1)$$

where $q = [q_{s_1}, q_{s_2}, q_{s_N}]$ denotes the labeled score distributions of the images, $\hat{q} = [\hat{q}_{s_1}, \hat{q}_{s_2}, \hat{q}_{s_N}]$ is predicted score distributions, $s = \{s_1, s_2, ...s_N\}$ denotes the ordered score, N is the total number of score buckets, $CDF_q(k)$ is the cumulative distribution function, r is set to 2 to penalize the Euclidean distance between the CDFs.

A. Vision Stream

Unlike previous works [18], [29], which use the random cropping strategy to extract multiple patches as inputs, we employ the recurrent attention convolutional network to iteratively extract visual features on important spatial regions. It consists of two components: the high-level feature extractor and the recurrent attention generator.

The high-level feature extractor can be based on any type of CNN, and we take VGG-16 as an example in this section. Each input image i_p is resized to 224×224 and then fed into the CNN to extract the deep features. In our model, we cut the normal VGG-16 and take the first 5-convolutional blocks. The output of the last convolutional block is a tensor with dimensions (W, H, D), where W and H represent the spatial resolution, and D is the channel dimension. Thus, each input image i_p can be represented by $W \times H$ vectors. Each vector is a D-dimensional representation corresponding to the local part of the image. Then, the output of the last convolutional block can be represented as follows:

$$F = \{f_1, f_2, ..., f_L\} \quad f_i \in R^D, \quad (2)$$

where F denotes the output feature of the pre-trained VGG-16, $L = W \times H$ is the number of locations, and f_i is the i-th feature vector in F.

After obtaining these high-level features, we can then focus on the key regions to generate image representations, as shown in Fig. 3. Recurrent attention is calculated via the long short-term memory (LSTM) network since it has a cell memory to determine what kind of information to forget and what kind of information to memorize. The detailed architecture of the LSTM neural network and the attention module are illustrated in Fig 4. The hidden layer in LSTM is computed as follows:

$$i_t = \sigma(W_{xi}x_t + W_{ci}c_{t-1} + W_{hi}h_{t-1} + b_i) \quad (3)$$
$$f_t = \sigma(W_{xf}x_t + W_{cf}c_{t-1} + W_{hf}h_{t-1} + b_f) \quad (4)$$
$$o_t = \sigma(W_{xo}x_t + W_{co}c_{t-1} + W_{ho}h_{t-1} + b_o) \quad (5)$$
$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (6)$$
$$h_t = o_t \tanh(c_t) \quad (7)$$

where i is the input gate, f is the forget gate, o is the output gate, c is the memory cell, x_t is the input features, σ is the logistic sigmoid function, W_* is the weight matrix and b_* is the bias vector. The hidden state h_{t-1} of the LSTM network can be incorporated with image features F to predict the attention map. The attention map is computed by a 2-layer feed-forward

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2985526, IEEE Transactions on Multimedia

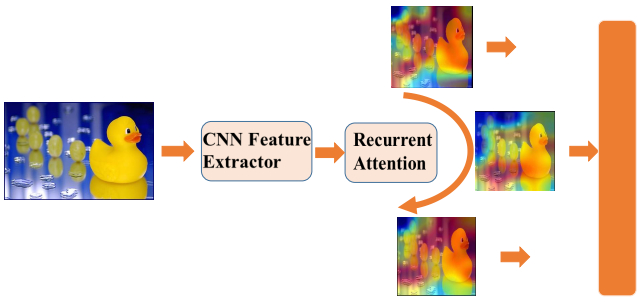


Fig. 3. The recurrent attention structure. LSTM produces attention maps in each iteration. These attentional representations are finally combined to form the final visual features.

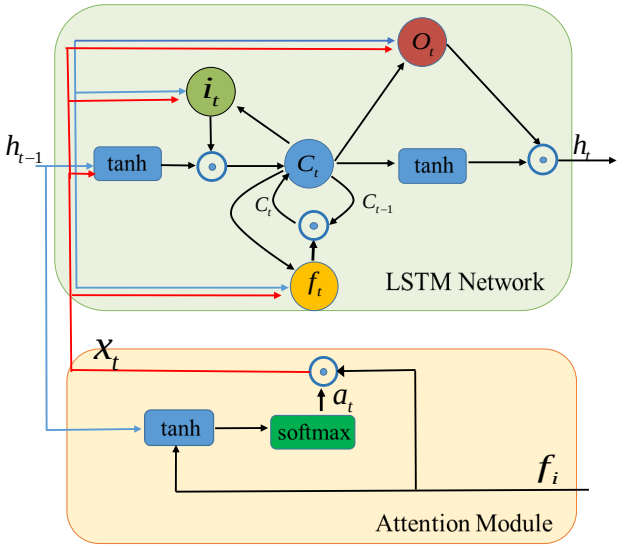


Fig. 4. The structure of the LSTM unit and the attention module. The blue arrows and red arrows represent the input of h_{t-1} and x_t, respectively.

neural network (FNN) and a softmax operator:

$$e_{t,i} = N_i tanh(W_1 h_{t-1} + W_2 f_i + b) \tag{8}$$

$$a_{t,i} = \frac{exp(e_{t,i})}{\sum_{k=1}^{L} exp(e_{t,k})}, \tag{9}$$

where N_i, W_1, and W_2 are the network parameters estimated during training. After obtaining the attention map, the new image features are calculated by multiplying the attention weights directly with the input image features F:

$$\begin{aligned} x_t &= \sum_{i=1}^{L} a_{t,i} f_i \\ a_t &= \{a_{t,1}, a_{t,2}, ..., a_{t,L}\}, \quad t \in 1, ..., T, \end{aligned} \tag{10}$$

where a_t is the attention mask matrix. The final visual features are the concatenation of these attention representations, *i.e.*, $x = [x_1, x_2, ..., x_T]$.

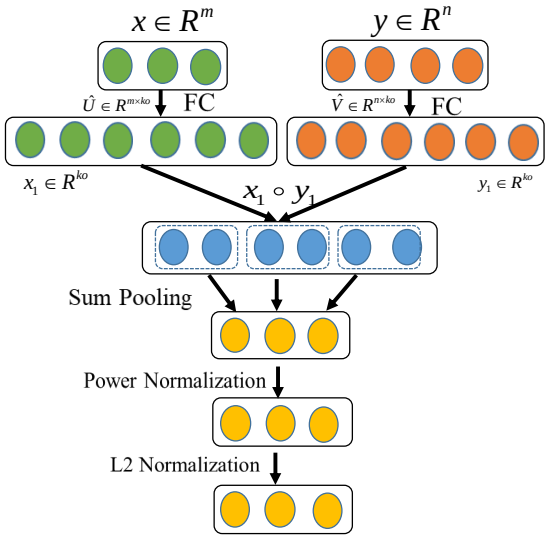


Fig. 5. The structure of the MFB pooling module.

B. Language Stream

The language features are extracted based on the Text-CNN [20] model. Different from CNN models in image classification, the input to the Text-CNN is sentences or documents represented by a matrix. Each row of the matrix represents a word (one token), and the vectors in each row are often the word embeddings from word2vec [42] or GloVe [43]. Let $d_i \in R^k$ be the k-dimensional word vector corresponding to the i-th word. These k-dimensional vectors are extracted from GloVe [43] in this paper. A sentence of length n words is represented as a $n \times k$ matrix. This is the input "image" to the Text-CNN. Then, the convolution operation is performed on these matrices:

$$y_i = f(wd_{i:i+h-1} + b). \tag{11}$$

where y_i is the output feature, f is a nonlinear function, $w \in R^{h \times k}$ is the convolutional filter, and b is a bias term. In vision, the filters slide over local patches of the image, but the convolutional filters of Text-CNN slide over full rows of the matrix. That is why the convolutional filters in Text-CNN have fixed size "width" k but have varying size "height" h. To capture multiple contextual relationships, we adopt filters that have three height sizes, e.g., $3, 4, 5$. Finally, a max-pooling operation over these feature maps is performed. These features are combined together for further processing.

C. Multimodal Factorized Bilinear Pooling

The main challenge of multimodal approaches lies in how to optimally fuse the multimodal features. Generally, concatenation or elementwise summations are most frequently used for multimodal feature fusion. However, features in different modalities often have different distributions; thus, they vary significantly. Simple linear fusion methods fail to encode the complex interactions between features. In this paper, we adopt

multimodal factorized bilinear pooling [23] to fuse the features from the vision stream and language stream.

Given the visual feature $x \in R^m$ and the textual feature $y \in R^n$, the multimodal bilinear models can be defined as follows:

$$s_i = x^T Z_i y \tag{12}$$

where s_i is the multimodal feature fusion output and $Z_i \in R^{m \times n}$ is the projection matrix that encodes the complex relationship between features. Inspired by the matrix factorization tricks, (12) can also be rewritten as follows:

$$\begin{aligned} s_i &= x^T U_i V_i^T y = \sum_{d=1}^{k} x^T u_d v_d^T y \\ &= SumPool(Ux \circ Vy, k) \end{aligned} \tag{13}$$

where $U_i = [u_1, u_2, ...u_k] \in R^{m \times k}$ and $V_i = [v_1, v_2, ...v_k] \in R^{n \times k}$ are the factorized matrices, k is the latent dimensionality of the factorized matrices, $\circ$ is the elementwise multiplication, and SumPool indicates using a k-size nonoverlapped window to perform sum pooling. The above analysis assumes that the output $s_i \in R$. In general, the output often has o-dimensions, i.e., $s \in R^o$. In this case, we need to learn the projection matrix $Z = [Z_1, Z_2, ..., Z_o]$. Then the factorized matrices U and V change into three-order tensors: $U = [U_1, ..., U_o] \in R^{m \times k \times o}$ and $V = [V_1, ..V_o] \in R^{n \times k \times o}$, respectively. We reformulate U and V into 2-D matrices: $\hat{U} \in R^{m \times ko}$ and $\hat{V} \in R^{n \times ko}$. Thus, (13) is rewritten as follows:

$$s = SumPool(\hat{U}^T x \circ \hat{V}^T y, k) \tag{14}$$

According to (13), MFB pooling can be easily implemented with some commonly used layers, such as a fully connected layer and pooling layers. To prevent overfitting, a dropout layer and normalization layers are also used. The detailed structure is illustrated in Fig. 5.

IV. Experiments

In this section, extensive experiments are conducted to verify the effectiveness of the proposed MRACNN. First, the dataset and evaluation metrics are introduced. Secondly, the implementation details are reported. Thirdly, ablation studies are performed to verify each component of MRACNN. Finally, the superiority of the proposed method is discussed in comparison with some state-of-the-art methods.

A. Dataset and Evaluation Criteria

AVA Dataset: In this paper, we adopt the largest publicly available aesthetics assessment dataset, *i.e.* the Aesthetic Visual Analysis (AVA) dataset [44] to train our model. The AVA dataset includes $250,000$ images that are collected from www.dpchallenge.com. Each image has approximately 200 voters, and the score ratings range from 1 to 10. We find that some users not only rate images but also leave comments to explain why they rated an image as such. These comments encode the viewpoints of users and provide additional information for aesthetic prediction. However, the AVA dataset only contains visual images. Thus, we enrich the AVA dataset by crawling all the user comments for images. Among the total $250,000$ images, there are $1,526$ images that have no user comments at all. Some images have useless comments that only contain one word or numbers. We removed these images and finally obtained $223,716$ images in total. We used $17,380$ images for testing and $206,336$ images for training. The partition of training data and testing data are the same as in previous work [12], [18], [29].

Photo.net Dataset: The Photo.net dataset is one of earliest datasets in the image aesthetic prediction area. It consists of $20,278$ images collected from https://www.photo.net/. Just like AVA dataset, it only contains visual images, we enrich the photo.net dataset by crawling all its user textual comments. Among the $20,278$ images, $13,583$ images have effective user comments. For a fair comparison, we use the same data partition as the previous work [30] in which $11,754$ images are used for training, 782 images are used for validation, and the rest $2,409$ images are used for the test.

Evaluation Criteria: To compare with traditional methods fairly, we evaluate our proposed method with respect to three aesthetic quality tasks: (i) aesthetic score regression, (ii) aesthetic quality classification, and (iii) aesthetic score distribution prediction. We first predict the distribution of ratings for a given image. Then, we can obtain the aesthetic regression results by computing the mean score of the distributions via $\mu = \sum_{i=1}^{N} s_i \times p_{s_i}$. Finally, the aesthetic quality classification results can be obtained by thresholding the mean score using the threshold 5 just as the work of [12], [24], [26], [29]. Images with predicted scores above 5 are categorized as high quality, and images with scores lower than 5 are treated as low-quality images. For the image aesthetic score regression, we use four performance measures: Spearman rank-order correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC), root mean square error (RMSE) and mean absolute error (MAE). Of these criteria, the SRCC and PLCC measures present monotonicity between the ground-truth subjective visual quality scores and their predicted values, while MAE and RMSE report the prediction error. For the image aesthetic quality classification task, we adopt the most widely used evaluation metric: the overall accuracy, which is defined as $Accuracy = \frac{TP+TN}{P+N}$. For the image aesthetic label distribution prediction, we use the EMD to measure the closeness of the predicted and ground-truth rating distribution with $r = 1$ in (1).

B. Training Details

In this section, we first explain the basic training parameter setting. Then, we emphatically discuss some important parameters and different network structures, *i.e.*, the number of recurrences, the number of LSTM layers, and the high-level feature extractors. For these parameters, we conducted several experiments to choose the most suitable one.

1) Basic Training Parameter Setting: For the language stream, we tokenize the review texts and then use 300 dimensional GloVe [43] to initialize the word embedding. The parameters of the embedding are fixed during training across all of our experiments. The filter windows in the Text-CNN

are 3, 4, 5 with 100 feature maps each. The dropout rate is set to 0.3. The maximum input length is set as 200 for each batch in the dataset. We adopt zero-padding for other sequences that are beneath this limit. For the vision stream, it consists of three parts: the convolutional feature extractor, LSTM network and attention module. The convolutional feature extractor and the parameters in LSTM are discussed in the following section. In all experiments, we use the Adam solver with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The base learning rate is 0.001 and reduced by a factor of 10 every 10 epochs. The batch size is set to 32. The training continues until the validation loss reaches a plateau for 5 epochs. Our networks are implemented based on the open source PyTorch framework with a NVIDIA Pascal TITAN X GPU.

2) Recurrence Number T: The number of recurrences T is the steps that the attention is performed. It is an important parameter since it controls the complexity of the model. In this paper, we tried three parameters, 1, 3, 5. More recurrence numbers can also be chosen, but it adds more computational burden. The experimental results are illustrated in Table I and Fig. 6. According to the experiments in Table I, the accuracy, SRCC, and PLCC slightly decrease when $T > 3$, while the EMD is almost saturated. The validation loss of $T = 5$ is not lower than the loss of $T = 3$ in Fig. 6(a), which suggests that no further improvements are possible when increasing the recurrence number. Thus, in this paper, we select $T = 3$.

3) Layer Number of LSTM: We set the layer number in LSTM 1, 2, and 3, respectively. Experimental results for this parameter are also shown in Table I and Fig. 6(b). As seen, a network that is too deep decreases the classification accuracy and increases the validation loss. Therefore, we set the LSTM layers as 1 in our network.

4) Convolutional Feature Extractor: Our proposed MARCNN can be incorporated with any type of network architecture. Therefore, we investigated several CNNs as our visual feature extractor. According to the depth of the network, the CNNs we selected are AlexNet [45], VGG-16 [46], and Inception-v3 [47]. AlexNet is relatively small compared with other networks. VGG-16 [46] is often used as a feature extractor by many traditional aesthetic prediction methods, such as [12], [26], [29], [36]. Inception-v3 is carefully designed to increase the depth and width of the network, but it has an auxiliary classifier on top of the last 17×17 layer. As a feature extractor, we remove the auxiliary classifier and keep the size of the output features as 17×17. All the above networks are initialized with models trained on ImageNet and fine-tuned on the AVA dataset. The comparison results are shown in Table II. As expected, the proposed MARCNN architecture outperforms the baselines by a large margin on all three networks.

C. Ablation Studies

We evaluate the contribution of each component in the MRACNN architecture. For this purpose, we construct six different variations: a single column network that is only based on visual features (vision stream), a single column network that is based on textual features (language stream), a multimodal fusion network without an attention mechanism (VGG16+Text-CNN), and three other baselines with different multimodal fusion models. All ablation study experiments are based on the VGG-16 network architecture. For a fair comparison, all the training parameters are unified.

Effectiveness of Multimodal Feature Fusion. To validate the effectiveness of the multimodal feature fusion, we compare it with two different variants. The first variant is "Text-CNN", which predicts the aesthetics only with the language stream. The other is "VGG-16", which predicts image aesthetics via the vision stream. "Text-CNN+VGG-16" predicts image aesthetics by combining vision and language streams. Table III reports the experimental results. From the table, we can make the following observations. First, the classification result of the language stream is promising. *Text-CNN* achieves 81.85% and 0.7786 in classification accuracy and SRCC, which is a 2.42% and 2.93% improvement compared with *VGG-16*. Second, combining vision and language boosts performance significantly. The *Text-CNN+VGG-16* achieves 84.94% in classification accuracy, while *Text-CNN* and *VGG-16* achieve 81.85% and 78.43%, respectively. This is a 3.09% and 6.51% performance improvement. For SRCC, the performance improvement is even larger, *i.e.,* 17.4% (compared to *VGG-16*) and 4.47% (compared to *Text-CNN*), respectively. This demonstrates that visual information and user comments are complementary in aesthetic prediction.

Effectiveness of Recurrent Attention. To iteratively focus on the important regions to extract visual information, we further add the recurrent attention network to the high-level feature extractor *VGG-16* and denote it as *MRACNN* in Table III. *Text-CNN+VGG-16* only achieves 84.94% in classification accuracy, 0.8233 in SRCC, but *MRACNN* achieves 85.68% and 0.8310, respectively. This illustrates that the recurrent attention mechanism can indeed improve the performance of aesthetic prediction tasks. Fig. 7 shows some examples of attention maps on the AVA dataset. We can clearly see that the learned attention focuses on the important regions in the image, which verifies our assumption.

Effectiveness of MFB Module. We then compare the performance of MFB with other multimodal fusion methods, *i.e.*, feature concatenation, elementwise summation, elementwise product. Fig. 8 shows the courses of validation. We can clearly see that MFB achieves the lowest validation loss, highest validation accuracy and SRCC during training. Table III shows the overall comparison results on the test sets. It can be seen that there is a constant improvement on all metrics when using the MFB module.

D. Comparison with State-of-the-Art Methods

In this section, we quantitatively compare with ten state-of-the-art methods to verify the superiority of our proposed approach. Five deep learning methods aim to perform binary classification, and one method aims to perform aesthetic regression, namely, *MTCNN* [26], *A-Lamp* [29], *MNA-CNN* [12], *RAPID* [19], *DMA-Net* [18], and *MLSP* [31]. There are also two multimodal feature fusion methods, *i.e.*, *multimodal DBM* [36] and *Multigap* [48]. While the other two methods, *NIMA* [33] and *GPF-CNN* [30] are designed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2985526, IEEE Transactions on Multimedia

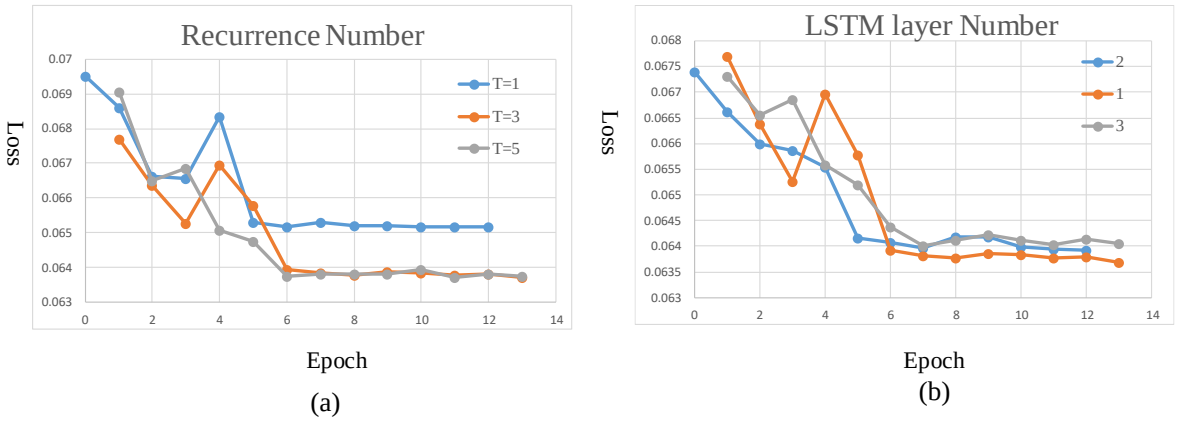


Fig. 6. Illustration of the loss on the validation set with different recurrence number and layer number. Panels (a) shows that no further improvements are possible when the recurrence number is larger than 3. Panels (b) shows that the network achieves the best performance when setting LSTM layer number to 1.

TABLE I
PERFORMANCE COMPARISON WITH DIFFERENT RECURRENCE NUMBERS AND NUMBER OF LSTM LAYERS

	Numbers	Accuracy (%)↑	SRCC(mean)↑	PLCC (mean)↑	MAE↓	RMSE ↓	EMD ↓
Recurrence Number	1	84.98	0.8223	0.8346	0.3359	0.4294	0.0378
	3	**85.68**	**0.8310**	**0.8431**	**0.3271**	**0.4196**	**0.0371**
	5	85.53	0.8308	0.8428	0.3278	0.4199	0.0371
Layer Number of LSTM	**1**	**85.68**	**0.8310**	**0.8431**	**0.3271**	**0.4196**	**0.0371**
	2	85.54	0.8292	0.8426	0.3294	0.4208	0.0372
	3	85.59	0.8286	0.8416	0.3298	0.4216	0.0373

TABLE II
PERFORMANCE OF DIFFERENT NETWORK ARCHITECTURE.

Network architecture	Accuracy (%)↑	SRCC(mean)↑	PLCC (mean)↑	MAE↓	RMSE ↓	EMD ↓
AlexNet	76.37	0.5549	0.5665	0.4733	0.6063	0.0525
VGG16	78.43	0.6493	0.6596	0.4299	0.5537	0.048
InceptionNet	79.43	0.6756	0.6865	0.4154	0.5359	0.0466
MRACNN(AlexNet)	**84.46**	**0.8202**	**0.8314**	**0.3377**	**0.4331**	**0.0381**
MRACNN(VGG-16)	**85.68**	**0.8310**	**0.8431**	**0.3271**	**0.4193**	**0.0371**
MRACNN(InceptionNet)	**85.71**	**0.8318**	**0.8434**	**0.3252**	**0.4190**	**0.0369**

TABLE III
ABLATION STUDY.

Network architecture	Accuracy (%)↑	SRCC(mean)↑	PLCC (mean)↑	MAE↓	RMSE ↓	EMD ↓
Text-CNN	81.85	0.7786	0.7851	0.3895	0.4887	0.0447
VGG16	78.43	0.6493	0.6596	0.4299	0.5537	0.048
Text-CNN+VGG16	84.94	0.8233	0.8367	0.3345	0.4274	0.0377
MRACNN(EltwiseProd)	85.26	0.8270	0.8403	0.3302	0.4229	0.0382
MRACNN(EltwiseSum)	85.36	0.8265	0.8399	0.3308	0.4214	0.0388
MRACNN(Concat)	85.32	0.8260	0.8388	0.3316	0.4247	0.0387
MRACNN(MFB)	**85.68**	**0.8310**	**0.8431**	**0.3271**	**0.4193**	**0.0371**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2985526, IEEE Transactions on Multimedia

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS ON AVA DATASET. V DENOTES THE VISUAL FEATURE AND T DENOTES THE TEXTUAL FEATURES.

Backbone	Network architecture	Accuracy (%)↑	SRCC(mean)↑	PLCC (mean)↑	MAE↓	RMSE ↓	EMD ↓
AlexNet	RAPID(V) [19]	74.2	-	-	-	-	-
AlexNet	DMA-Net(V) [18]	75.42	-	-	-	-	-
VGG16	MNA-CNN(V) [12]	76.1	-	-	-	-	-
VGG16	Multi-DBM(V+T) [36]	78.88	-	-	-	-	-
VGG16	A-Lamp(V) [29]	82.5	-	-	-	-	-
VGG16	MTCNN(V) [26]	78.46	-	-	-	-	-
VGG16	GPF-CNN(V) [30]	80.70	0.6762	0.6868	0.4144	0.5347	0.046
VGG16	NIMA(V) [33]	80.6	0.592	0.610	-	-	0.052
InceptionNet	NIMA(V) [33]	81.51	0.612	0.636	-	-	0.05
InceptionNet	Multigap(V+T) [48]	82.27	-	-	-	-	-
InceptionNet	MLSP(V) [31]	81.72	0.756	0.757	-	-	-
InceptionNet	GPF-CNN(V) [30]	81.81	0.6900	0.7042	0.4072	0.5246	0.045
AlexNet	**MRACNN(V+T)**	**84.46**	**0.8202**	**0.8314**	**0.3377**	**0.4331**	**0.0381**
VGG-16	**MRACNN(V+T)**	**85.68**	**0.8310**	**0.8431**	**0.3271**	**0.4193**	**0.0371**
InceptionNet	**MRACNN(V+T)**	**85.71**	0.8318	**0.8434**	**0.3252**	**0.4190**	**0.0369**

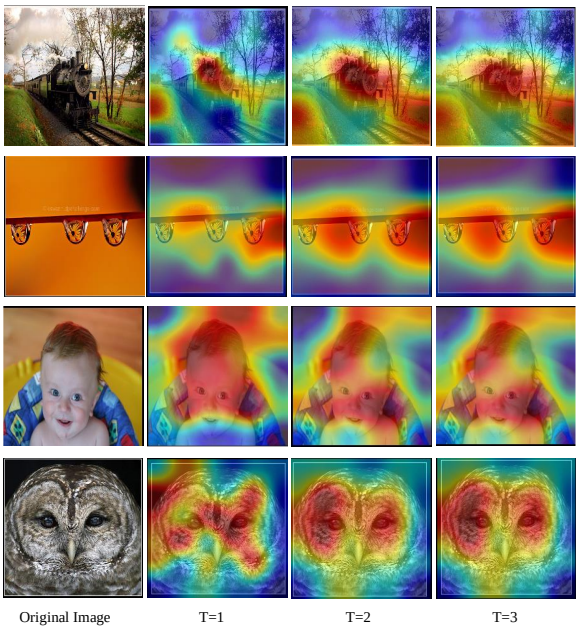


Fig. 7. Examples of attention maps on the AVA dataset. T is the recurrence number.

to perform unified aesthetic prediction on three tasks, which are quite similar to us. Among all ten baselines, some [18], [19] are based on AlexNet [45], others [26], [29], [33], [36] are based on VGG-16 [46], or InceptionNet [47]. For a fair comparison, we report the results of *MRACNN* on those three networks: *i.e.* AlexNet, VGG-16 and InceptionNet. Table IV shows the comparison results. *RAPID* and *DMA-Net* are based on shallow AlexNet, achieving 74.2% and 75.42% in classification accuracy, respectively. However, our *MRACNN*(AlexNet) achieves 84.46%. This is 10.26% and 9.03% improvements, respectively. Other networks such as *MNA-CNN*, *MTRLCNN*, and *A-Lamp* are based on the larger deep network VGG-16. Among them, *A-Lamp* achieves the highest accuracy rate, i.e., 82.5%. Our *MRACNN*(VGG-16) obtains 85.68%, which is a 3.81% point improvement. The *multimodal DBM model* and *Multigap* [48] are closely related to ours since they aim to fuse the textual and visual features to predict the aesthetics. *Multigap* extends *multimodal DBM model* to achieve 82.27% accuracy by using InceptionNet to extract the features. In contrast, our MRACNN (InceptionNet) achieves 85.71% accuracy, which shows the superiority of our method. *MLSP* [31] extracts multilevel features from all convolutional blocks of a pre-trained InceptionNet and integrates these features to predict image aesthetics. Our method overcomes it with an improvement of 3.99% according to the overall accuracy metric and 7.58% according to the SRCC metric. *NIMA* and *GPF-CNN* both aim to predict the score distributions of the image. *GPF-CNN*(InceptionNet) uses neural attention to guide the network to extract fine details and thus outperforms *NIMA*, achieving 81.81% in classification accuracy and 0.69 in SRCC. However, our *MRACNN*(InceptionNet) achieves the strongest results at 85.71% accuracy rate and 0.8318 in SRCC. This is 3.89% better than *GPF-CNN* and 4.19% better than *NIMA* according to the overall accuracy metric. For the SRCC metric, this is a notable 14.18% improvement compared with *GPF-CNN* and 24.98% improvement compared with *NIMA*. To the best of our knowledge, this is the best performance on the standard AVA test set. Fig. 9 shows some example images predicted by our MRACNN model. The top row shows some predicted high-quality images, and the bottom row illustrates some low-quality images. The prediction results can almost perfectly reconstruct the ground-truth label in some cases.

E. Content-based Photo Aesthetic Analysis

In this section, we conduct comprehensive experiments to test the effectiveness of MRACNN on various types of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2985526, IEEE Transactions on Multimedia

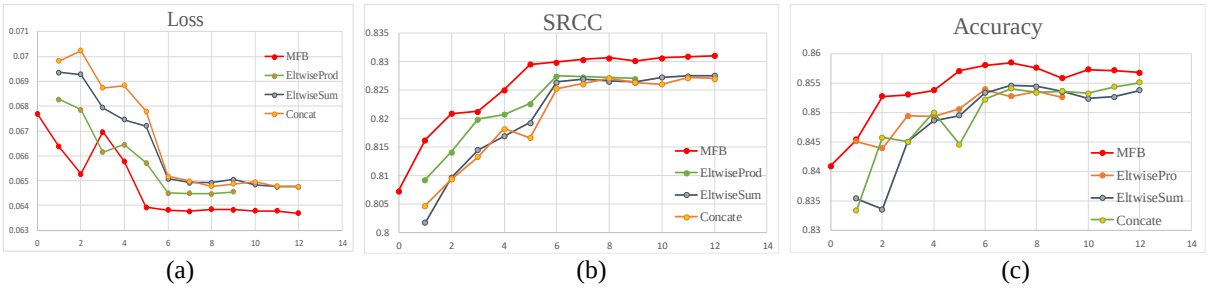


Fig. 8. Performance comparison of MFB and other baseline multimodal fusion models. (a) Comparison of Validation loss versus epoch. (b) Comparison of SRCC versus epoch. (c) Comparison of validation accuracy versus epoch. We can easily find that MFB achieves the lowest validation loss, highest validation accuracy and SRCC during training.

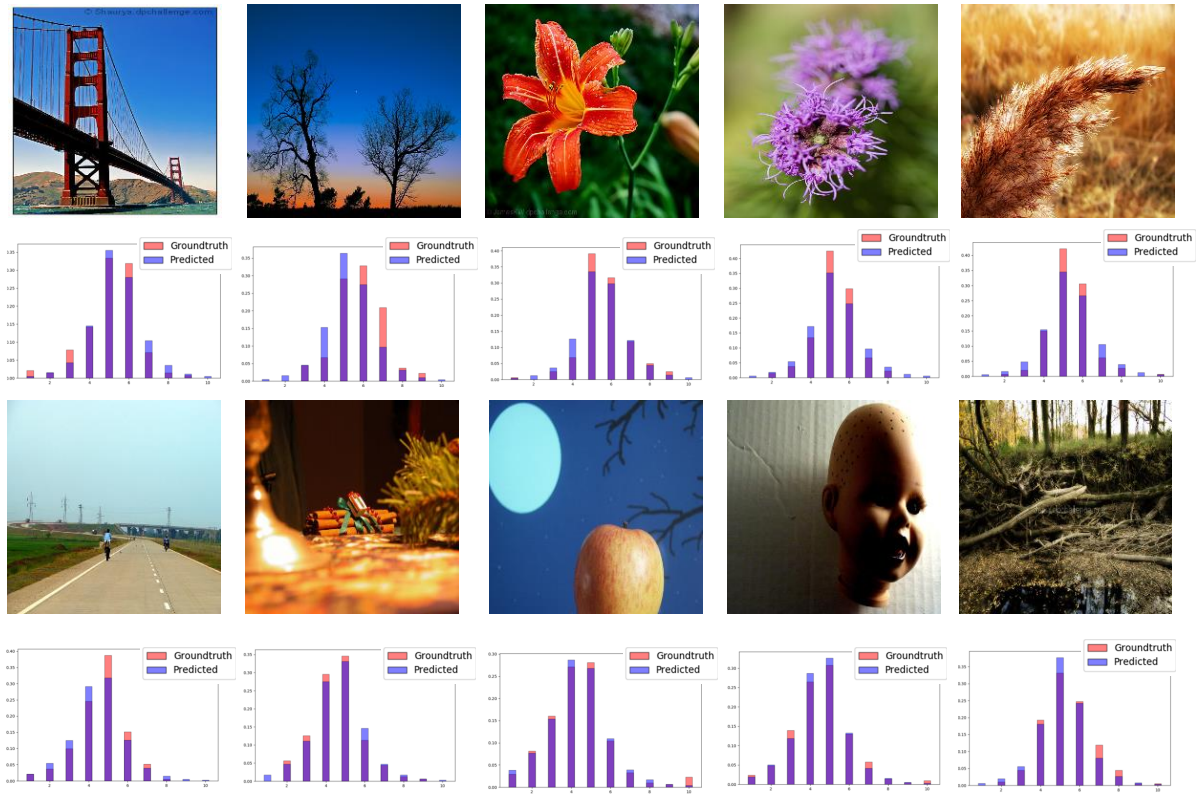


Fig. 9. Randomly selected examples from the AVA dataset. Top two rows: high-quality images, as predicted by our MRACNN (VGG16), coupled with plots of their ground-truth and predicted score distributions. Bottom two rows: the low-quality images, as predicted by our MRACNN (VGG16), coupled with plots of their ground-truth and predicted score distributions.

images. Following the work of [12], [29], [44], we select eight categories: *i.e.* animal, landscape, cityscape, floral, food-drink, architecture, portrait, still life from the AVA test sets. Each category consists of $2,200$ images, and each image has at least one user comment. For all eight categories, we systematically compared *MRACNN* with Text-CNN and two other state-of-the-art methods, i.e., *NIMA* [33] and *GPF-CNN* [30]. For the *NIMA*, we implement it with the VGG-16 network and use the EMD loss to train the network. The comparison results are shown in Table V. Specifically, the MRACNN achieves the best performance in most of the categories, where animal and still life show substantial improvements compared with *NIMA* and *GPF-CNN*. We note that still life is the most difficult category to classify. However, MRACNN achieves 78.08% in the classification accuracy rate, which is a 6.94% relative improvement over *NIMA*, and a 1.73% improvement compared with *GPF-CNN* , confirming the effectiveness of our proposed method.

F. Evaluating Performance on Photo.net Dataset

To further evaluate the performance of the proposed *MRACNN*, we compare it with several methods, ranging from

TABLE V
ABLATION STUDY ON EIGHT CATEGORY IMAGES.

Category	Network architecture	Accuracy (%)↑	SRCC(mean)↑	PLCC (mean)↑	MAE↓	RMSE ↓	EMD ↓
animal	**MRACNN**	**82.57**	**0.8295**	**0.8400**	**0.3235**	**0.4150**	**0.0375**
	Text-CNN	78.79	0.7724	0.7887	0.3754	0.4734	0.0444
	GPF-CNN [30]	80.80	0.7480	0.7478	0.3941	0.5051	0.045
	NIMA(VGG16) [33]	76.17	0.6212	0.6267	0.4959	0.6325	0.0546
landscape	**MRACNN**	**86.43**	**0.8325**	**0.8503**	**0.3165**	**0.4004**	**0.0359**
	Text-CNN	82.74	0.7753	0.7940	0.3725	0.4718	0.0448
	GPF-CNN [30]	85.42	0.7746	0.7822	0.3713	0.4705	0.0422
	NIMA(VGG16) [33]	80.04	0.6780	0.6876	0.4968	0.6276	0.0541
cityscape	**MRACNN**	81.41	**0.8156**	**0.8304**	**0.3403**	**0.4338**	**0.0382**
	Text-CNN	78.17	0.7609	0.7819	0.3898	0.4902	0.0451
	GPF-CNN [30]	**81.68**	0.7533	0.7539	0.3956	0.5103	0.0443
	NIMA(VGG16) [33]	76.22	0.6460	0.6424	0.5074	0.6481	0.0552
floral	**MRACNN**	**80.33**	**0.7997**	**0.8142**	**0.3235**	**0.4114**	**0.0368**
	Text-CNN	77.68	0.7402	0.7640	0.3598	0.4571	0.0422
	GPF-CNN [30]	79.95	0.7374	0.7348	0.3681	0.4785	0.0423
	NIMA(VGG16) [33]	75.58	0.6184	0.6220	0.4455	0.5709	0.05
fooddrink	**MRACNN**	**81.85**	**0.8204**	**0.8359**	**0.3225**	**0.4107**	**0.0362**
	Text-CNN	80.99	0.7786	0.7851	0.3895	0.4887	0.0447
	GPF-CNN [30]	80.22	0.7389	0.7476	0.3919	0.4948	0.0443
	NIMA(VGG16) [33]	74.01	0.6163	0.6278	0.4876	0.6208	0.0536
architecture	**MRACNN**	**82.22**	**0.8091**	**0.8301**	**0.3171**	**0.3988**	**0.0358**
	Text-CNN	79.65	0.7494	0.7729	0.3612	0.4563	0.0422
	GPF-CNN [30]	81.60	0.7431	0.7410	0.3704	0.4822	0.0421
	NIMA(VGG16) [33]	76.83	0.6032	0.6069	0.4748	0.6091	0.0521
portrait	**MRACNN**	**84.14**	**0.8142**	**0.8268**	**0.3348**	**0.4289**	**0.0381**
	Text-CNN	81.77	0.7714	0.7787	0.3814	0.4824	0.0449
	GPF-CNN [30]	83.52	0.6987	0.7047	0.4228	0.5389	0.0475
	NIMA(VGG16) [33]	76.93	0.5671	0.5726	0.5347	0.6784	0.0583
still life	**MRACNN**	**78.08**	**0.7852**	**0.8057**	**0.3445**	**0.4392**	**0.0385**
	Text-CNN	73.58	0.7254	0.7506	0.3906	0.4948	0.0441
	GPF-CNN [30]	76.35	0.7001	0.7127	0.4039	0.5153	0.0455
	NIMA(VGG16) [33]	71.14	0.5652	0.5810	0.49	0.6187	0.0537

TABLE VI
COMPARISON WITH STATE-OF-THE-ART METHODS ON PHOTO.NET DATASET.

Network architecture	Accuracy (%)↑	SRCC(mean)↑	LCC (mean)↑	MAE↓	RMSE ↓	EMD ↓
GIST_SVM [11]	59.90	-	-	-	-	-
FV_SIFT_SVM [11]	60.8	-	-	-	-	-
MTCNN(VGG16) [26]	65.2	-	-	-	-	-
GPF-CNN(VGG16) [30]	75.6	0.5217	0.5464	0.4242	0.5211	0.070
MRACNN(VGG-16)	**78.91**	**0.5709**	**0.5902**	**0.3636**	**0.4589**	**0.0622**

handcrafted methods, such as GIST and FisherVector [11], to deep learning methods *MTCNN* [26] and *GPF-CNN* [30]. Considering that the compared methods are based on the VGG-16 network, we choose the VGG-16 as the backbone network to achieve a fair comparison. Table VI illustrates the results. From the table, we find that this dataset is much more challenging than the AVA dataset because the same method usually obtains lower accuracy. It needs to be noted that our *MRACNN* still keeps the best and achieves 78.91% for aesthetic classification accuracy and 0.5709 on the SRCC metric. It also reduces the EMD distance to 0.4589. To the best of our knowledge, this is the most comprehensive performance evaluation reported on the Photo.net dataset in the literature.

V. Conclusion

In this study, we presented a method of deep multimodality learning for image aesthetic prediction tasks. We constructed the AVA comment dataset and the photo.net comment dataset. These datasets can advance the research on multimodal modelling in image aesthetics. Additionally, we designed a novel two stream network to make full use of the complementary information between user comments and images. The first stream employed the recurrent attention network which can iteratively focus on the important regions to extract the visual features. The second stream used the Text-CNN to encode the semantic features from user comments. Finally, the Multimodal Factorized Bilinear (MFB) pooling is used to achieve effective feature fusion. The effectiveness of each component has been validated through an extensive evaluation. We also demonstrate that the proposed method outperforms state-of-the-art methods by a large margin.

Although our method achieves promising results, several aspects need to be considered in our future research. First, our model assumes that the multimodal data is present both during training and inference steps. Thus it is not able to handle missing multimodal data. Future research can potentially address this problem in several ways. The first possible method is to reconstruct the multimodal representation via probabilistic graphical models, which allows for missing data. Another possibility is to generate textual descriptions related to aesthetics. Second, visual and textual information collaboratively describe photo aesthetic. The intrinsic correlation between these two features needs to be further investigated. Third, most of existing methods are based on deep learning networks which lack interpretability. It is imperative to design explainable aesthetic prediction models, which not only output aesthetic scores but also generate a textual description explaining why the image have a high aesthetics score.

References

[1] F.-L. Zhang, M. Wang, and S.-M. Hu, "Aesthetic image enhancement by dependence-aware object recomposition," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1480–1490, 2013.

[2] A. Samii, R. Měch, and Z. Lin, "Data-driven automatic cropping using semantic composition search," *Computer Graphics Forum*, vol. 34, no. 1, pp. 141–151, 2015.

[3] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proceedings of the ACM International Conference on Multimedia*, 2010, pp. 271–280.

[4] F. Gao, J. Yu, S. Zhu, Q. Huang, and Q. Tian, "Blind image quality prediction by exploiting multi-level deep representations," *Pattern Recognition*, vol. 81, pp. 432–442, 2018.

[5] Y. Liu, K. Gu, S. Wang, D. Zhao, and W. Gao, "Blind quality assessment of camera images based on low-level and high-level statistical features," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 135–146, 2018.

[6] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proceedings of European Conference on Computer Vision*, 2008, pp. 386–399.

[7] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[8] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proceedings of European Conference on Computer Vision*, 2006, pp. 288–301.

[9] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1657–1664.

[10] H. Su, T. Chen, C. Kao, W. H. Hsu, and S. Chien, "Scenic photo quality assessment with bag of aesthetics-preserving features," in *Proceedings of the International Conference on Multimedia*, 2011, pp. 1213–1216.

[11] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proceedings of IEEE International Conference on Computer Vision*, 2011, pp. 1784–1791.

[12] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 497–506.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015.

[14] K. He, X. Zhang, and S. Ren, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] X. Tian, Z. Dong, K. Yang, and T. Mei, "Query-dependent aesthetic model with deep learning for photo quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2035–2048, 2015.

[16] S. Kong, X. Shen, Z. L. Lin, R. Mech, and C. C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proceedings of 14th European Conference on Computer Vision*, 2016, pp. 662–679.

[17] B. Jin, M. V. O. Segovia, and S. Süsstrunk, "Image aesthetic predictors based on weighted cnns," in *Proceedings of International Conference on Image Processing*, 2016, pp. 2291–2295.

[18] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 990–998.

[19] X. Lu, Z. L. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.

[20] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014.

[21] A. Hu and S. R. Flaxman, "Multimodal sentiment analysis to explore the structure of emotions," in *Proceedings of the ACM International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 350–358.

[22] X. Bai, M. Yang, P. Lyu, Y. Xu, and J. Luo, "Integrating scene text and visual appearance for fine-grained image classification," *IEEE Access*, vol. 6, pp. 66 322–66 335, 2018.

[23] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 29, no. 12, pp. 5947–5959, 2018.

[24] L. Guo, Y. Xiong, Q. Huang, and X. Li, "Image esthetic assessment using both hand-crafting and semantic features," *Neurocomputing*, vol. 143, pp. 14–26, 2014.

[25] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 33–40.

[26] Y. Kao, R. He, and K. Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Trans. Image Processing*, vol. 26, no. 3, pp. 1482–1495, 2017.

[27] M. Kucer, A. C. Loui, and D. W. Messinger, "Leveraging expert feature knowledge for predicting image aesthetics," *IEEE Trans. Image Processing*, vol. 27, no. 10, pp. 5100–5112, 2018.

[28] Y. Chen, Y. Hu, L. Zhang, P. Li, and C. Zhang, "Engineering deep representations for modeling aesthetic perception," *IEEE Trans. Cybernetics*, vol. 48, no. 11, pp. 3092–3104, 2018.

[29] S. Ma, J. Liu, and C. W. Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, July 21-26, 2017, pp. 722–731.

[30] X. Zhang, X. Gao, W. Lu, and L. He, "A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction," *IEEE Trans. Multimedia*, vol. 21, pp. 2815–2826, 2019.

[31] V. Hosu, B. Goldlucke, and D. Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9375–9383.

[32] N. Murray and A. Gordo, "A deep architecture for unified aesthetic prediction," 2017. [Online]. Available: https://arxiv.org/abs/1708.04890

[33] H. Talebi and P. Milanfar, "NIMA: neural image assessment," *IEEE Trans. Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[34] X. Jin, L. Wu, X. Li, S. Chen, S. Peng, J. Chi, S. Ge, C. Song, and G. Zhao, "Predicting aesthetic score distribution through cumulative jensen-shannon divergence," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 77–84.

[35] X. He and Y. Peng, "Fine-grained image classification via combining vision and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7332–7340.

[36] Y. Zhou, X. Lu, J. Zhang, and J. Z. Wang, "Joint image and text representation for aesthetics analysis," in *Proceedings of the ACM Conference on Multimedia Conference*, 2016, pp. 262–266.

[37] G. Wang, J. Yan, and Z. Qin, "Collaborative and attentive learning for personalized image aesthetic assessment," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2018, pp. 957–963.

[38] A. M. Obeso, J. Benois-Pineau, M. S. García-Vázquez, and A. A. Ramírez-Acosta, "Forward-backward visual saliency propagation in deep nns vs internal attentional mechanisms," in *Proceedings of the International Conference on Image Processing Theory, Tools and Applications*, 2019, pp. 1–6.

[39] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proceedings of the Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 2204–2212.

[40] X. Sun, H. Yao, R. Ji, and S. Liu, "Photo assessment based on computational visual attention model," in *Proceedings of the 17th ACM international conference on Multimedia*, 2009, pp. 541–544.

[41] K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang, and B.-G. Hu, "Attention-based multi-patch aggregation for image aesthetic assessment," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 879–886.

[42] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the Neural Information Processing Systems.*, 2013, pp. 3111–3119.

[43] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.

[44] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2408–2415.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, 2012, pp. 1106–1114.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

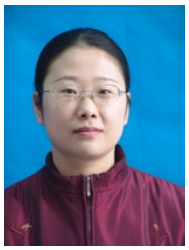
[48] Y.-L. Hii, J. See, M. Kairanbay, and L.-K. Wong, "Multigap: Multi-pooled inception network with text augmentation for aesthetic prediction of photographs," in *Proceedings of IEEE International Conference on Image Processing*, 2017, pp. 1722–1726.

Xiaodan Zhang received the B.Eng. degree in electronic information engineering and Ph.D. degree in intelligence information processing from Xidian University, Xi'an, China, in 2014 and 2019. Her current research interests include image aesthetic quality assessment, visual attention, and computationally modeling of human visual system.

Xinbo Gao (M'02-SM'07) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. He is currently a Cheung Kong Professor of Ministry of Education of P. R. China, a Professor of Pattern Recognition and Intelligent System, and the Dean of Graduate School of Xidian University. His current research interests include Image processing, computer vision, multimedia analysis, machine learning and pattern recognition. He has published six books and around 300 technical articles in refereed journals and proceedings. Prof. Gao is on the Editorial Boards of several journals, including Signal Processing (Elsevier) and Neurocomputing (Elsevier). He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences. He is a Fellow of the Institute of Engineering and Technology and a Fellow of the Chinese Institute of Electronics.

Wen Lu received the BSc, MSc and PhD degrees in signal and information processing from Xidian University, China, in 2002, 2006and 2009 respectively. He is currently the professor at Xidian University and postdoctoral research in the department of electrical engineering at Stanford University, USA. His research interests include image quality metric, human vision system, computational vision. He has published 2 books and around 30 technical articles in refereed journals and proceedings including IEEE TIP, TSMC, Neurocomputing, Signal processing etc.

Lihuo He is currently an associate professor at Xidian University. He received the B.Sc.degree in electronic and information engineering and Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2008 and 2013. His research interests focus on image/video quality assessment, cognitive computing, and computational vision.



Jie Li received the B.Eng. degree in electronic engineering, the M.Sc. degree in signal and information processing, and the Ph.D. degree in circuit and systems from Xidian University, Xi'an, China, in 1995, 1998, and 2004, respectively. She is currently a Professor with the School of Electronic Engineering, Xidian University. She has published around 50 technical articles in refereed journals and proceedings including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and Information Sciences. Her current research interests include image processing and machine learning.