IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Deep Neural Framework with Visual Attention and Global Context for Predicting Image Aesthetics

YIFEI XU[1,4], NUO ZHANG[1], PINGPING WEI[2], GENAN SANG[3], LI LI[3], and FENG YUAN[3]

[1]School of software, Xi'an Jiaotong University, 710054, Xi'an, Shaanxi, China. (e-mail: belonxu_1@xjtu.edu.cn; zn1111@stu.xjtu.edu.cn)
[2]State Key Laboratory for Manufacturing Systems Engineering, Xi'an jiaotong University, 710054, Xi'an, China (e-mail: erin1989@xjtu.edu.cn)
[3]Alltuu Inc., 801, Block A, Bld. 3, No.1508, Wenyi West Road, 311100, Hangzhou, Zhejiang, China (e-mail: sanggenan@alltuu.com; yuanfeng@alltuu.com;lili@alltuu.com)
[4]Huiyichen Inc. 1703, Block 1, No 1388, Jiulonghu Ave, 330038, Nanchang, Jiangxi, China

Corresponding author: YIFEI XU (e-mail: belonxu_1@xjtu.edu.cn).

This research was funded by part by the Young Scientists Fund of the National Natural Science Foundation of China under Grant 61802300, China Postdoctoral Science Foundation Funded Project under grant 2018m643666, Xi'an jiaotong university basic research foundation for Young Teachers under grant xjh012019043, and National Science and Technology Major Project under grant 2019YFB2102501 and 2019YFB2103005.

ABSTRACT Computational inference of aesthetics has recently become a hot topic due to its usefulness in widely applications such as evaluating image quality, retouching image and retrieving image. Owing to the subjectivity of this problem, there is no general framework to predict image aesthetics. In this paper, we propose a deep neural framework with visual attention module, self-generated global features and hybrid loss to address this problem. Specifically, the framework can be any state-of-the-art convolution classification network compatible with visual attention. Further, self-generated global feature compensates for the loss of global context information during training stage, and the hybrid loss guides the network to learn the similarity between the predicted aesthetic scores and the ground-truths through fusing soft-max-entropy and Earth Mover's Distance(EMD). With the above-mentioned improvements, the proposed deep neural framework is capable of effectively predicting image aesthetics in an efficient way. In our experiments, we release a real-world aesthetic dataset that contains 1,800 2K photos labeled by several experienced photographers, and then provide a thorough ablation study of the design choices to better understand the superiority brought by each part of our framework, and design several comparisons with the state-of-the-art methods on a fraction of metrics. The experimental results on two datasets demonstrate that both accuracy and efficiency achieve favorably performance.

INDEX TERMS Image Aesthetics, Hybrid Loss, Deep Neural, Visual Attention

I. INTRODUCTION

IMAGE aesthetic quality assessment (IAQA) is a long-standing visual task, which lays foundation in many multi-media applications such as image retouching, image ranking and image retrieval. Practically, photographic retouchers use photograph editing software to enhance images based on human aesthetic quality. As the pre-processing step, they need to quickly make objective assessments of numerous images. However, the above procedure is time-consuming and intractable in real-world application. Therefore, it is essential to design an outperforming model to assess image aesthetic quality quickly.

The goal of IAQA is to design the algorithms which automatically predict image aesthetic quality. Definitely, it is challenging as the aesthetic score of given images relies on several undetermined factors, such as composition, color distribution, technical quality and so on. To address the problem, earlier approaches aim to classify aesthetic attributes of an image using hand-crafted features and achieve good progress. However, hand-crafted features depend heavily on expert knowledge, and cannot capture feature presentations comprehensively.

To overcome this shortcoming, deep neutral models such as Convolutional Neural Network (CNN) were proposed recently [1]–[8], which related image content to semantic level qualities, and extracted a lot of aesthetic features without human interaction. Now that noticeable benefit has been

made by deep neural network, they are still faced with the following limits: 1) Most of deep neural models focus on classifying images to low/high score or regressing to the mean score, but our goal is to predict IAQA with higher correlation with human ratings. 2) The CNN-based deep neural models never consider global features that could reflect global lighting condition or even subject types. 3) Current deep neural models ignore the attention mechanism that can explore a focused location and suppress unnecessary ones. 4) Loss function is not well suitable to describe the training loss when classification labels are in order.

To address the preceding limits, this paper presents a new deep IAQA framework for predicting image quality. In comparison with traditional binary IAQA classification, we devote to predicting the distribution of human ratings. Considering that image aesthetic assessment is affected by global features, we propose and incorporate global features into our proposed network. If IAQA method could focus on important features and depressing unnecessary one, it is beneficial to achieve more objective results in the task of IAQA. Consequently, we introduce and integrate visual attention module with our proposed network to further boost the performance. In order to learn score distribution and score values of the samples simultaneously, we also proposes a hybrid loss composed of Earth Mover's Distance(EMD) loss and soft-max cross-entropy. Using datasets AVA and Alltuu, our experimental results reveal the superiority of our framework against the state-of-the-art.

To sum up, the major contributions of our work can be listed as follows. 1) We propose a deep IAQA framework by embedding CNN-based classifier with self-generated global features and attention module to improve the performance. 2) We release a new dataset of 1,800 2K HD photos, each of which is labeled by several expert photographers. 3) We perform the evaluation of our framework on different datasets, and demonstrate its superiority qualitatively and quantitatively.

II. RELATION WORK

Existing IAQA methods can be roughly divided into two kinds. Methods of the first kind try to map hand-crafted features into image aesthetic score with a shadow architecture. Methods of the second kind rely on impressive progress on deep neural networks for IAQA [9]. In the rest of this section, a few representative works are briefly discussed.

A. TRADITIONAL MODELS

Inspired by image processing, perception and photography, traditional models first extract features based on well-designed or general guidelines, and then appended a classifier to distinguish image quality. Early algorithms use specified descriptor, such as edge distribution and rule of the third, to create the features for classifier or regression. [10] attempts to imitate how humans perceive photo aesthetic quality based on the spatial distribution of edges. Later, [11], [12] compute both global and local compositional features to discover object regions, and present saliency-enhanced methods for the classification of professional photos and snapshots. Even recently, researchers still construct well-designed image aesthetic attributes by sharpness, depth, charity, tone and color [13]. Michal Kucer et at [14] combine hand-crafted feature and CNN-based features to predict image quality. Several methods adopt general descriptors such as scale-invariant features transform (SIFT) descriptor or local patches of colors to describe image. [15] evaluates the color harmony of a collection of local regions from a given image, and uses SVM to determine whether the image has a high or low aesthetic quality. In [16], Marchesotti et al. use the bag of visual words and fisher vector to achieve impressive performance. Bhattacharya et al. [17] present an interactive application that enables users to improve visual aesthetics of their digital photographs using spatial recomposition, and provides an optimal location of each foreground object for the users. Although the models based on traditional features are effective in some particular datasets or applications, they heavily depend on the expert experience and are less descriptive to represent the complex objects in different domains.

B. DEEP NEURAL MODELS FOR IAQA

Recently, deep neural models has been developed to solve several classical problems, such as fault diagnosis and fault tolerant control [18], [19], object detection [20], image classification [2], [21] and so on. Success of CNN on object classification task provides a new perspective on IAQA. Recent works in the literature can be divided into two major schemes: 1) Deep neural models based on rank numerical score. 2) Deep neural models based on score distribution. Methods of the first scheme extract deep aesthetic features, and attach them to the rank score with a canonical classifier. In earlier works, researchers direct adopt generic deep features learned from other tasks to train a new classier. In order to represent aesthetic feature better, various CNN-based models are trained from scratch directly with single-column CNNs [22]–[25]. In detail, these models add skip connection or replace convolutional layers to explore the potential in learning the aesthetic presentation, and then concatenate the output features with fully connected layers. Instead of focusing on single-column network, researchers pay more attention to multi-column CNNs. Lu et al. [26] attempt to tackle the aesthetic modeling problem through the two-column network called RAPID which captures global features and local features from these two columns, respectively. Later, to address the limitations of global layout encoder and fixed-sized input, References [27], [28] aggregate the two network layers and add an object-based attributed graph. Despite the excellent performance, the above methods only focus on the connection between the output predictions with numerical score ranking, but ignore the real demand. Recently, researchers turn attention to the deep neural models considering score distribution (i.e., a score distribution vector of the ordinal basic human ratings). Roy et al. [29] predict image aesthetics by using inferential information depending

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/ACCESS.2020.3015060, IEEE Access

on visual content found in an image. Talebi et al. [30] propose NIMA, an approach that calculates aesthetic scores from predicted aesthetics rating distributions. Further, as an extension of NIMA, Wang et al. [3] utilize aspect-ratio-preserving multi-patch learning approach for predicting aesthetic scores. However, additional computation and regions of the selected patches stunt its promotion in practical application. Moreover, some recent works [1], [5] formulate IAQA as an obviously visible high or low binary classification problem, and report promising classification accuracy on benchmark. Although it is impractical and infeasible to use binary classifier in real-world application, we still consider them just for comparison.

III. THE PROPOSED FRAMEWORK

In this section, we present our general training and testing framework to assess image scores. Inspired by [30], the framework could perform well based on image classifiers.

A. THE OVERALL FRAMEWORK FLOW

The pipeline of the proposed framework is shown in Fig.1, which includes the CNN-based classifier with attention module, self-generated global feature module (SGFM) and hybrid loss function. In detail, the classifier could be any state-of-the-art CNN-based network, such as VGG16 [31], Inception-V2 [32], Inception-ResnetV2 [33], DenseNet [34] and so on. Similar to NIMA, we also replace the last layer of the classifier with a fully-connected layer followed by a soft-max activations. Rather than processing a whole scene at once, we incorporate attention module into the classifier to selectively focus on salient parts. As for SGFM, it could reveal high-level information which may be helpful for individual pixels to determine their local contribution [35], so we decide to encode implicit global features for classification task. During training procedure, the image with its corresponding score vector is fed into classifier network. The goal of our framework is to minimize the hybrid loss between the predicted probabilities vector and the ground-truths. During testing procedure, the mean and variance deviation operation are applied to predict aesthetic score distribution. To sum up, the learning procedure are listed in Algorithm 1.

B. THE VISUAL ATTENTION MODULE

The visual attention mechanism has been proved effective in expressing human perception. As feasible and plug-and-play modules, SENET block [36] and CBAM block [37] are cleverly designed to compute attention in different aspects and verified in different datasets (MS-COCO dataset and VOC dataset). Thus, we incorporate these attention modules into the classifiers of our framework to improve the accuracy performance. Here, a brief introduction of the two modules is listed as below.

1) Squeeze-and-Excitation Network (SENET)

SENET introduces a compact model to explore channel relationship by feature recalibration technology, through which the network can learn global average-pooled features to selectively emphasis informative features and suppress less useful ones. In detail, SENET contains two steps: squeeze and excitation. The goal of squeeze is to exploit channel dependencies through squeezing global spatial information into a simple channel descriptor. The squeeze operation is formulated as 1.

$$\mathbf{z}_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{1}$$

where the output $\mathbf{z} \in \mathbb{R}^C$ is generated by squeezing $\mathbf{U}$ through spatial dimensions $H \times W$, $\mathbf{C}$ is the channel dimension of the given feature map $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_C]$. In fact, we can find that the transformation of $\mathbf{U}$ can be regarded as a collection of expressive statistics for each channel. Excitation operation aims to make full use of the squeeze aggregated information. To implement it, as Fig.2 shows, a gating mechanism with a sigmoid activation is adopted to represent the weights of channels.

$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \mu(\mathbf{W}_2 \delta(\mathbf{W}_1)\mathbf{z}) \tag{2}$$

where μ denotes ReLu function, $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$, $\mathbf{W}_1$, and $\mathbf{W}_2$ are the weights of dimensionality-reduction layer with reduction ratio *r* and dimensionality-increasing layer, respectively. The final output is computed by multiplying the feature map $\mathbf{U}$ with scaler S_c.

$$\mathbf{x}_c = s_c \cdot \mathbf{u}_c \tag{3}$$

2) Convolutional Block Attention Module(CBAM)

To overcome the limitation of SENET [37], CBAM leverages attention ability in the following points: 1) Using average-pooled feature and max-pooled feature simultaneously when computing channel-wise attention. 2) Appending a spatial attention sub-module to channel attention sub-module. Compared with SENET, CBAM is superior with a little more computational burden. The CBAM module is a computational unit which can be embedded in any feature map. Let $\mathbf{U} \in \mathbb{R}^{W \times H \times C}$ denotes the feature map, 1D channel attention map $\mathbf{M_c} \in \mathbb{R}^{1 \times 1 \times C}$ and 2D spatial attention map $\mathbf{M_s} \in \mathbb{R}^{H \times W \times 1}$ are applied to $\mathbf{U}$ sequentially. Formally, the overall attention process can be defined as:

$$\begin{aligned} \mathbf{U}^{'} &= \mathbf{M_c}(\mathbf{U}) \otimes \mathbf{U} \\ \mathbf{U}^{''} &= \mathbf{M_s}(\mathbf{U})^{'} \otimes \mathbf{U}^{'} \end{aligned} \tag{4}$$

where $\otimes$ is element-wise multiplication, $\mathbf{U}^{'}$ and $\mathbf{U}^{''}$ are the outputs of channel attention and spatial attention, respectively. Firstly, CBAM produces a channel attention map by exploring the inter-channel relationship of feature happened in SENET. Beyond the canonical idea of aggregating spatial information, CBAM uses both average-pooled and max-pooled features to infer fine channel-wise attention. Then, CBAM generates a spatial attention through utilizing inter-spatial relationship of features. To compute spatial attention, average-pooling and max-pooling operations along the

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/ACCESS.2020.3015060, IEEE Access

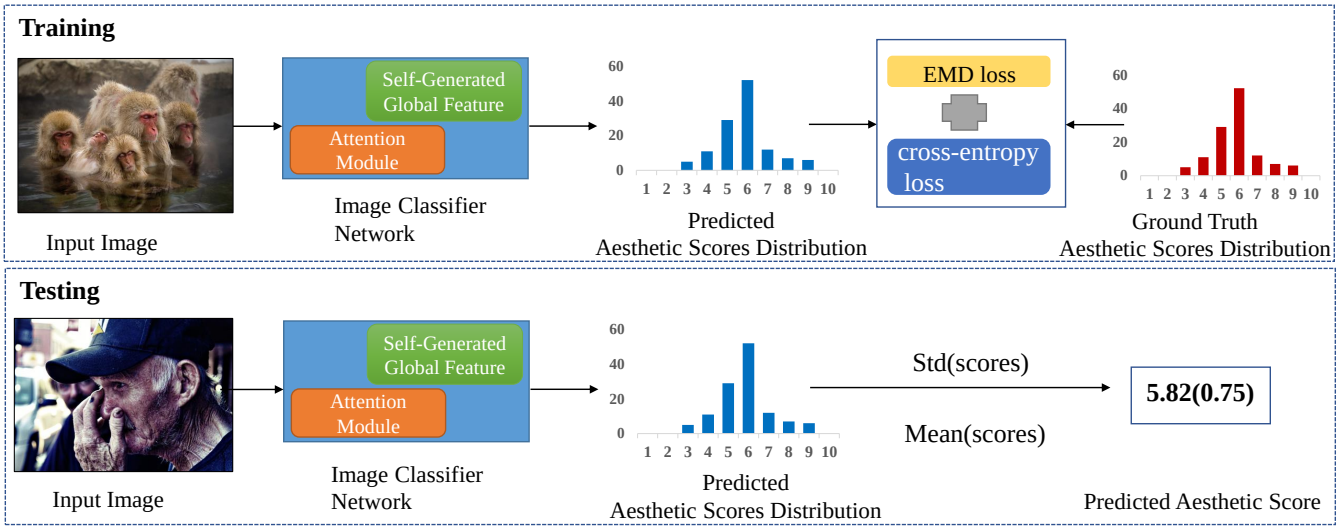

FIGURE 1. The pipeline of the proposed framework.

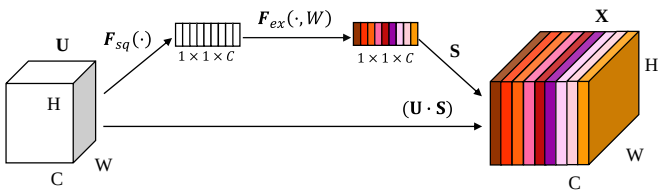

FIGURE 2. The flowchart of SENET.

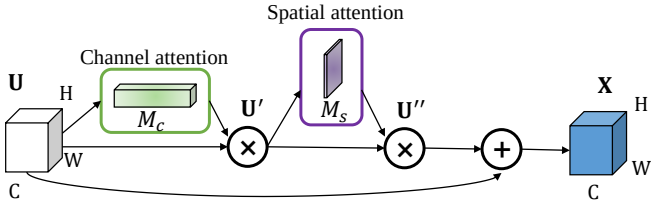

FIGURE 3. The flow of CBAM, where ⊕ is element-wise addition.

channel axis are concatenated to generate informative feature descriptor. Finally, as shown in Fig.3, CBAM employs a sequential manner with channel-first and be integrated with a given CNN-based Network.

C. SELF-GENERATED GLOBAL FEATURE MODULE (SGFM)

Global features have been explored by other image processing tasks for long history. However, their models require extra supervised network trained with explicit scene labels. Besides, the residual learning has been shown helpful on convergence for image processing tasks. Therefore, motivated by these two ideas, we define an implicit global feature module based on the network itself, and named it self-generated global feature model (SGFM). To be concrete, the module contains two contraction layers and other components that go with them. Each contraction layers contains 5×5 filters with stride 2 followed by SELU activation and batch normalization [38], and other components consist of fully-connected layer, SELU activation, reduction layer and a few related operations.

Formally, the general idea of SGFM can be expressed as follows: Firstly, denoting the feature map $M \times M \times N$ of the classifier, we reduce it to $M/2 \times M/2 \times N$ and then $M/4 \times M/4 \times N$ filtered by the aforementioned contraction layer twice. Secondly, the feature map $M/4 \times M/4 \times N$ is squeezed into $1 \times 1 \times N$ by virtue of a fully-connected layer followed by a SELU activation and then another fully-connected layer. Then, the $1 \times 1 \times N$ feature map is duplicated $M \times M$ copies and then concatenated with the original $M \times M \times N$ feature map, resulting a $M \times M \times 2N$ concatenated feature map. Finally, it is reduced by reduction-G and restored to the size $M \times M \times N$. Empirically, we give some instructions how to incorporate SGFM into existing CNN Networks. Since SGFM tends to describe global information of an given image, it is more appropriate to apply it as early as possible. If it is performed before the softmax layer, abundant critical global information are not learned well by the CNN network. Consequently, we decide to put SGFM after the fist reduction layer or convolution block to preserve global information as much as possible. Specially, we integrate SGFM with two state-of-the-art classifiers: Inception-Resnet-V2 and densenet169. Fig.4 illustrates the basic network architecture of modified Inception-Resnet-V2 (named Inception-Resnet-V2(attention&G)). As can be seen, given an input image with size of $299 \times 299 \times 3$, after processed by Stem block, the network would output $35 \times 35 \times 256$ feature maps ($M = 35, N = 256$), and feed them into SGFM to capture informative features. As demonstrated in the right panel of Fig.4, attention module are embedded

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/ACCESS.2020.3015060, IEEE Access

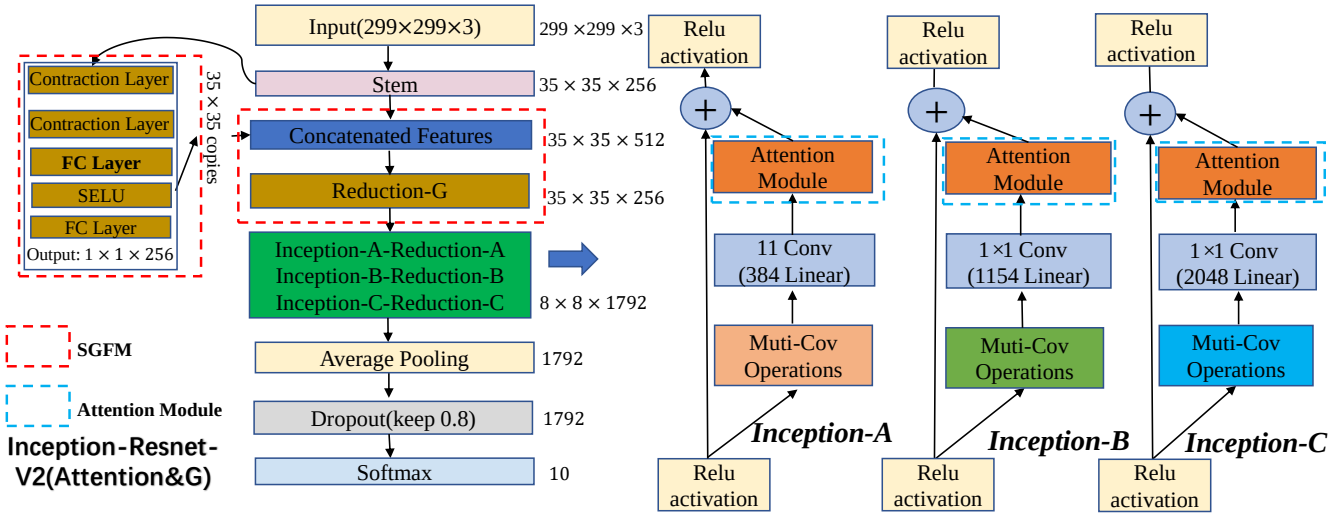

FIGURE 4. Inception-Resnet-V2 with SGFM and attention module.

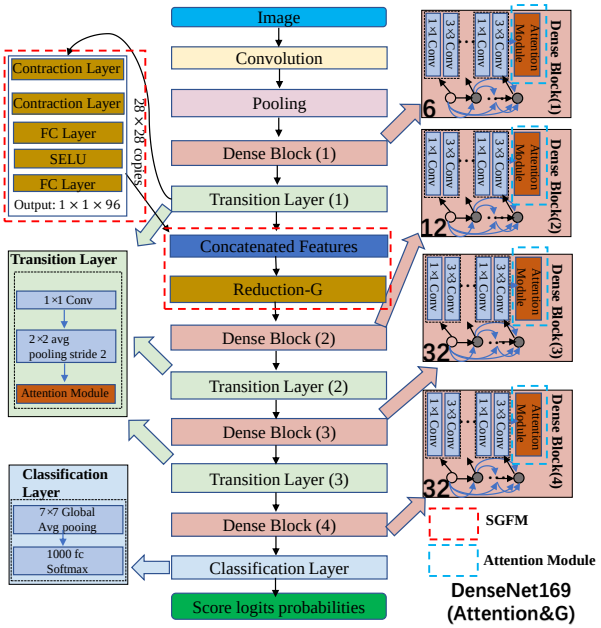

FIGURE 5. DenseNet with global feature module and attention module.

into Inception A-C at the branch of residual connection, respectively. Definitely, the attention module could be either CBAM block or SENET block. Different from Inception and Resnet, DenseNet connects each layer to every layer in a feed-forward way and makes a progressive achievement on various datasets (CIFAR-100, ImageNet, and etc.). In Fig.5, we empirically append SGFM to Transition Layer (1) and incorporate attention model into each Dense Block and Transition Layer.

D. HYBRID LOSS FUNCTION

In this work, we propose a new hybrid loss for IAQA, which contains EMD loss and softmax cross-entropy. Softmax cross-entropy loss can be written as $\sum_{i=1}^{N} -p_{s_i} \log(\hat{p}_{s_i})$ where $\hat{p}_{s_i}$ and p_{s_i} are the estimated probability and ground-truth label of *i*-th score bucket, respectively. Softmax cross-entropy is good at describing the local value of each class and prove helpful in classification tasks. However, softmax cross-entropy loss performs not well in the case of ordered-classes. To overcome the limit, EMD loss minimizes the cost to move the mass of estimated probability distribution to ground-truth probability distribution. For *N*-class aesthetic ratings, the value of the *i*-th rating class p_{s_i} is *i*, where $1 \leq i \leq N$. The *r*-norm distance of *i*-th rating class and *j*-th rating class is defined as $|s_i - s_j|_r$. In that case, as shown in [30], *r*-norm EMD loss between the above-mentioned rating distributions is computed as follow:

$$\mathrm{EMD}^r = \left(\sum_{k=1}^{N} |\mathrm{CDF}_{\mathbf{p}}(k) - \mathrm{CDF}_{\hat{\mathbf{p}}}(k)|^r \right)^{\frac{1}{r}} \tag{5}$$

where $\mathrm{CDF}_{\mathbf{p}}(k)$ and $\mathrm{CDF}_{\hat{\mathbf{p}}}(k)$ denote the cumulative distribution function of the ground-truth rating distribution $\mathbf{p}$ and the predicted rating distribution $\hat{\mathbf{p}}$, respectively. Herein, *r* is specified as 2 as well as NIMA.

Particularly, the final mean score is computed in Algorithm 1. Softmax cross-entropy tends to make the final result closer to some integer value ranging from 1 to 10, and EMD loss emphasizes on the comparison of prediction distribution and ground-truth distribution. Thus, the combination of these two losses is able to cope with different types of datasets, such as Alltuu and AVA. In order to describe their relationships clear, we design a new hybrid loss to train our models to present score distribution and score value simultaneously.

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/ACCESS.2020.3015060, IEEE Access

Mathematically, hybrid loss can be expressed as:

$$L = \alpha L_s + \beta L_{emd} \tag{6}$$

where L_s and L_{emd} are softmax cross-entropy loss and EMD loss, respectively, α and β are their corresponding weights, and α+β=1.

Algorithm 1 The algorithm of our proposed framework for IAQA

Input:

The original image $\mathbf{I}$, max-epochs $\mathbf{E}$, the classifier with SGFM and attention module $C_{lassifier}$, The number of aesthetic ratings N. Test images T.

Output:

The predicted image score probability $\tilde{\mathbf{S}}$, mean score $\tilde{V}$.

Training Stage:

Initialize the network weights, learning rate, batch size, and other parameters;

for $t = 1; i < E; t++$ **do**

Train network by optimizing the softmax cross-entropy loss L_s;

Train network by optimizing the EMD loss L_{EMD};

Compute the hybrid loss $L = \alpha L_s + \beta L_{emd}$ and update the model weight parameters;

end for

Testing Stage:

Feed T into $C_{lassifier}$, and then output the $\tilde{\mathbf{S}}$;

Compute $\tilde{V} = \sum_{i=1}^{N} i \times \tilde{\mathbf{S}}_i$;

return $C_{lassifier}$, $\tilde{\mathbf{S}}$, and $\tilde{V}$.

IV. EXPERIMENTS AND MATERIALS

In this section, we detail the datasets for training and testing in Section IV-A. All the comparative experiments are carried out on our workstation, which is equipped with an Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, an NVIDIA TITAN V with 12 GB memory, 64GB RAM, and 1T SSD. All the comparative algorithms are developed using Tensorflow 1.4.

A. DATASETS

1) AVA dataset

We use the benchmark dataset for Aesthetic Visual Analysis (AVA) [39] to evaluate the proposed framework. It comprises 250,000 images collected from the online photography community website www.dpchallenge.com. Each image is associated with 10 stages of ratings, ranging from 1 to 10. The number of raters assigned to each image ranges from 78 to 649, and the average value is 210. Samples of the AVA dataset, mean score (mean square variance), normalized score histograms, and sample images are displayed in Fig.6. The AVA dataset is split into training set (230,000 images) and testing set (20,000 images).

2) Alltuu dataset:

Alltuu dataset was collected by ourself with the help of five professional retouchers who have been worked in the field of photography more than ten years. After independent labeling work finished, these five experts discussed and reached consensus on the optimal score for each image. It contains 1,800 images collected from our own massive image library. Similar to AVA dataset, each image was marked as integer numerical score value ranging from 1 to 10, and the score distribution was encoded in one-hot way. In Fig.7, a few examples with scores and score histogram of the overall dataset are illustrated. Actually, we can easily find that the distribution of scores is closely approximated to Gaussian distribution. Different from AVA dataset, only one integer score is kept, and this unconventional design strategy can be explained as below. 1) The ineluctable side effect of inadequate retouchers. In AVA dataset, the images are rated by the countless photographers around the world and the side effect of outliers can be eliminated through average operation. Nevertheless, in our own dataset, there is no enough experienced retouchers to do it. 2) The requirement of our practical application. In our real-world application, we hope the novices could learn experience about how to select photo from the experts as soon as possible, and assign a relatively objective integer quality score during photo selection. Hence, we have reasons to believe that our own dataset is constructed reasonably. To our best knowledge, it is the first aesthetic real-world dataset with 2K photos, which would stimulate the development of IAQA.

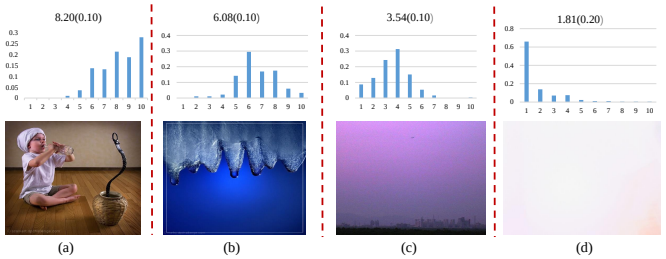

FIGURE 6. Mean score of sample image (mean square variance), normalized score histograms and sample images.

FIGURE 7. (a) The score histogram of Alltuu dataset, (b-d) A few samples of Alltuu dataset.

B. EXPERIMENTAL SETUPS

Before making a thorough comparison with various IAQA models, it is essential to determine the training parameters. First, for the comparative models, all layers apart from the last fully-connected layer are initialized by the parameters pre-trained on the ImageNet dataset [40]. The reason why we resort to pre-train technique is the model could efficiently

obtain feature extraction ability directly and greatly shorten training time. Then, all the images from these two datasets are resized to 324×324, random cropping with size 299×299 and random horizontal operations are applied to augment the datasets. During training stage, max-epochs $\boldsymbol{E}$ and learning rate is changed to 100 and 2×10^{-3}, respectively. Moreover, a momentum Adam optimizer (momentum=0.9) and a decay factor of 0.95 after every 10 epochs are used.

C. EVALUATION METRICS

We employ common-used metrics to quantitatively evaluate the performance of the comparative methods, ie. floating-point operations per second (Flops), the number of parameters of model (Params), mean absolute error(MAE), Pearson linear correlation coefficient (PLCC), spearman's rankorder correlation coefficient (SROCC) and mean squared error (MSE). To be fair, we report an extra metric called accuracy (ACC) for binary aesthetic quality classification. For the task of binary aesthetic quality classification, the images with average score higher than a threshold of $5+\sigma$ are deemed as positive examples and the rest are labeled as negative examples. Just to make it clear, the above-mentioned metrics are categorized into precision metrics and cost metrics depending on their properties. Precision metrics reflect the precision of models (MAE, PLCC, SROCC, MSE and ACC) and cost metrics represent the complexity and cost of models (Flops and Params). To sum up, lower Flops and Params indicate lighter computational costs, lower MAE, MSE values and greater PLCC, SROCC and ACC values signify higher precision.

V. RESULTS AND DISCUSSION

In this section, we conduct our experiments to evaluate the performance of our proposed methods for IAQA task. The experiments include two parts: One part evaluates our methods D169(C&G&H) and IRV(C&G&H) against several state-of-the-art methods on datasets Alltuu and AVA. It first respectively presents the overall comparison in multi-class aesthetic quality classification and binary aesthetic quality classification. Then, it compares the parameters and costs of all the methods to verify their efficiency. Finally, it depicts visual comparison on our dataset to verify the effectiveness of our visual attention module. The other part makes ablation studies of our framework. It demonstrates the different performance brought by different combinations of components appeared in our methods.

A. OVERALL COMPARATIVE STUDY

1) Multi-class Aesthetic Quality classification

In this section, we mainly compare our framework with previous state-of-the-art models. In order to verify the efficiency of our framework, two classical CNN-based classifier are chosen as baseline networks: Inception-Resnet-V2 and Denset169. The reason we prefer these two approaches is that they provide good baseline results either in terms of classification accuracy or computational consumption. Evaluation of our proposed methods with different architectures on datasets Alltuu and AVA are displayed in Table 1 and Table 2, respectively. We calculate Flops, Params, MAE, PLCC, MSE and SROCC metrics for comparative methods. Note that the implementation details of ASPP FCN-GC and SDLA are not released publicly, we cannot apply them directly in our comparation. In following discussion, we will not take them into account for multi-class aesthetic quality classification. Limited by discriminative ability of Inception V2, NIMA(IV) falls far behind the others on dataset Alltuu. In comparison with NIMA(IV), our IRV(C&G&H) reduces MAE and MSE by 0.095 (36.64%) and 0.104 (31.42%), and improve PLCC of aesthetics score by 0.032 (3.44%) and SROCC by 0.075(8.84%). Even compared with MPEMD, the performance of our IRV(C&G&H) gains significantly on precision metrics. For benchmark dataset AVA, compared with the previous best result obtained by MPEMD, PLCC shows an slight improvement of 0.019 (2.24%) and SROCC presents an improvement of 0.016 (2.33%). In contrast to NIMA(IV), PLCC shows an improvement of 0.075 (11.5%) and SROCC shows an improvement of 0.091 (14.87%). Now, let's take a close look at the results of models based on DenseNet169. Our model D169 (C&G&H) greatly surpasses NIMA(D169) on all precision metrics. When comparing with D169(C&G&H) and IRV(C&G&H), we observe that IRV(C&G&H) performs better on all precision metrics owning to the excellent baseline network. Consequently, we could come to an conclusion that the methods equipped with CBAM, SGFM and hybrid loss function outperform their comparatives on datasets Alltuu and AVA.

In addition, we further explore the consistency between the predicted aesthetic scores and the ground-truth with scatter plots on dataset Alltuu. As shown in Fig. 8, each point corresponds to an give image sample, the horizontal coordinate denotes the ground-truth, and the vertical coordinate denotes the predicted scores. Clearly then, most points locate around the diagonal line, indicating that most predicted scores are close to the corresponding ground-truth ones. From the above discussion, we can safely come to an conclusion that the combination of CBAM, SGFM and hybrid loss function play an important role in the proposed image aesthetic prediction framework.

2) Binary Aesthetic Quality Classification

To be fair, we also make a comparison for binary aesthetic quality classification task, and place the results on the rightmost column of Table1 and Table 2. During the comparative study, the source codes of ASPP FCN-GC and SDLA are not released and some experimental details are not mentioned. In this way, we only place their results reported in their papers. For purpose of comparing to existing binary classification results reported on dataset AVA, we simply set the threshold $\sigma = 0$ as the others do. Owing to the end-to-end forward architecture based on fully connection network, ASPP FCN-GC leverages the mutual dependencies

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/ACCESS.2020.3015060, IEEE Access

TABLE 1. Comparisons of the proposed framework with other state-of-the-art methods on dataset Alltuu. The rows above the first dashline present the results of Inception-V2 and its variants, and the bottom rows between the first dashline and the last dashline list the results of Denset169 and its variants. The rows below the last dashline are other comparative methods. For each metric, the best value is shown in bold.

Methods	MAE↓	PLCC↑	MSE↓	SROCC↑	ACC↑
NIMA(IV) [30][1]	0.257	0.909	0.435	0.848	90.9%
NIMA(IRV)[1]	0.202	0.922	0.372	0.876	93.4%
IRV(SENET)	0.190	0.927	0.351	0.889	94.6%
IRV(C)[1]	0.182	0.931	0.348	0.893	95.2%
IRV(G)[1]	0.191	0.925	0.353	0.887	94.3%
IRV(S)[1]	0.198	0.924	0.369	0.880	93.9%
IRV(H)[1]	0.190	0.928	0.351	0.890	94.9%
IRV(C&G)	0.179	0.934	0.340	0.910	96.4%
IRV(C&H)	0.175	0.937	0.337	0.912	96.1%
IRV(H&G)	0.182	0.932	0.346	0.902	95.5%
IRV(C&G&H)	**0.162**	**0.941**	**0.331**	**0.923**	97.0%
NIMA(D169)[1]	0.235	0.918	0.402	0.855	92.4%
D169(SENET)	0.217	0.921	0.376	0.862	93.4%
D169(C)	0.210	0.928	0.370	0.863	94.3%
D169(G)	0.222	0.920	0.387	0.860	92.7%
D169(S)	0.231	0.919	0.392	0.859	92.6%
D169(H)	0.210	0.921	0.372	0.862	94.0%
D169(C&G)	0.182	0.934	0.346	0.887	94.9%
D169(C&H)	0.179	0.937	0.344	0.882	95.5%
D169(H&G)	0.199	0.931	0.369	0.880	94.6%
D169(C&G&H)	**0.175**	**0.939**	**0.340**	**0.907**	96.1%
MPEMD [3]	0.181	0.935	0.344	0.883	89.5%
ASPP FCN-GC [5]	/	/	/	/	**97.3%**
SDLA [1]	/	/	/	/	96.7%

[1] IV:Inception-V2, IRV:Inception-Resnet-V2, IRV(C):Inception-Resnet-V2(CBAM), IRV(G):Inception-Resnet-V2(SGFM), IRV(H):Inception-Resnet-V2(Hybrid_loss), IRV(S):Inception-Resnet-V2(Softmax cross-entropy), D169:(DenseNet169)

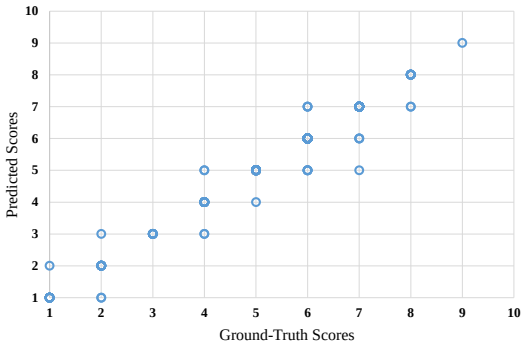

FIGURE 8. Scatter plots of the predicted scores vs. the ground-truth ones on Alltuu dataset

to boost aesthetic assessment and achieves the best performance on metric ACC. In spite of IRV(C&G&H) yields good results for multi-class aesthetic quality classification task measured by precision metrics, no improvement is shown for binary aesthetic quality classification. Meanwhile, MPEMD also performs not well in binary aesthetic quality classification as the optimization of EMD loss is more fit for multi-class classification than binary one. SDLA, a semi-supervised deep active learning algorithm, is good at discovering semantical perception of images assigned with contaminated tags. However, constricted by semi-supervised learning scheme, it performs less worse than ASPP FCN-GC. In Table1 and Table 2, thought ASPP FCN-GC works a little bit better than IRV(C&G&H), it need extra considerable computational cost. Consequently, we empirically observe that IRV(C&G&H) is comparable to ASPP FCN-GC to some extent. Additionally, we contribute a thorough discussion about the comparative results of MPEMD [3], NIMA(IV) [30], NIMA(IRV), IRV(C&G&H) and IRV(C&G&H). Using aspect-ratio-preserving multi-patch learning, aesthetic scores are obtained by predicting normalized aesthetics scores distribution. However, as shown in Table1 and Table 2, in conformity with work [3], MPEMD falls behind NIMA (IV) and IRV(C&G&H) for aesthetics binary quality classification. In work [3], lower ACC is contributed to the prediction bias around classification threshold. The difference between NIMA(IV) and NIMA(IRV) is the baseline network choice. Thanks to the higher discriminative capability of Inception-Resnet-V2, NIMA(IRV) performs better in binary classification. Further, SGFM, attention and hybrid loss are integrated with NIMA(IRV) to be our proposed IRV(C&G&H). Seen from Table1 and Table 2, IRV(C&G&H) boosts the performance marginally on metric ACC. Furthermore, we give a brief discussion on the performance of methods related to DenseNet169. As illustrated in Table 1 and Table 2, we still find the same conclusion that the choice of baseline network, SGFM, CBAM and hybrid loss play critical part in binary aesthetic quality classification.

3) The Overhead of Parameters and Computation

It is necessary to analyze the overhead of parameters and computation of IAQA models. Generally, shallower models are much more efficient than the deeper on cost metrics. Evidently, as show in Table 3, NIMA(IV) is significantly lighter than other models, but this comes at the cost of a apparent performance fall analyzed in above discussion. It is worthy noticing that D169(C&G&H) balances the trade-off of performance and computational cost, which can be a good choice when computational resource is limited. Particularly, SDLA works on the basis of probabilistic model, and no parameters and flops are reported. As for ASPP FCN-GC, it is too expensive to execute in real-world application even though it achieves the best performance in aesthetic binary classification. Thus, it is observed that perhaps D169(C&G&H) is the best choice when considering the influences between various factors.

4) Qualitative Comparison

When it comes to qualitative analysis, we employ GRAM-CAM [41] to visualize the compared networks on datasets Alltuu and AVA. GRAM-CAM is a recently proposed tech-

TABLE 2. Comparisons of the proposed framework with other state-of-the-art methods on dataset AVA. The rows above the first dashline present the results of Inception-V2 and its variants, and the bottom rows between the first dashline and the last dashline list the results of Denset169 and its variants. The rows below the last dashline are other comparative methods. For each metric, the best value is shown in bold.

Methods	MAE↓	PLCC↑	MSE↓	SROCC↑	ACC↑
NIMA(IV) [30][1]	0.280	0.636	0.321	0.612	81.51%
NIMA(IRV)[1]	0.252	0.692	0.290	0.686	81.73%
IRV(SENET)	0.243	0.693	0.280	0.688	81.97%
IRV(C)[1]	0.238	0.695	0.277	0.689	82.18%
IRV(G)[1]	0.249	0.693	0.282	0.687	81.82%
IRV(S)[1]	0.255	0.689	0.293	0.683	81.71%
IRV(H)[1]	0.242	0.694	0.277	0.688	82.15%
IRV(C&G)	0.232	0.696	0.274	0.691	82.34%
IRV(C&H)	0.229	0.697	0.272	0.692	82.48%
IRV(H&G)	0.237	0.695	0.276	0.689	82.23%
IRV(C&G&H)	**0.219**	**0.711**	**0.262**	**0.703**	83.52%
NIMA(D169)[1]	0.256	0.691	0.284	0.679	81.65%
D169(SENET)	0.250	0.692	0.282	0.682	81.74%
D169(C)	0.246	0.692	0.277	0.686	81.87%
D169(G)	0.253	0.691	0.283	0.680	81.70%
D169(S)	0.258	0.689	0.286	0.674	81.62%
D169(H)	0.247	0.692	0.280	0.685	81.79%
D169(C&G)	0.239	0.694	0.274	0.688	82.07%
D169(C&H)	0.233	0.695	0.272	0.690	82.29%
D169(H&G)	0.242	0.693	0.277	0.688	81.90%
D169(C&G&H)	**0.231**	**0.696**	**0.270**	**0.693**	83.11%
MPEMD [3]	0.233	0.692	0.276	0.687	79.38%
ASPP FCN-GC [5]	/	/	/	/	**83.59%**
SDLA [1]	/	/	/	/	83.09%

[1] IV:Inception-V2, IRV:Inception-Resnet-V2, IRV(C):Inception-Resnet-V2(CBAM), IRV(G):Inception-Resnet-V2(SGFM), IRV(H):Inception-Resnet-V2(Hybrid_loss), IRV(S):Inception-Resnet-V2(Softmax cross-entropy), D169:(DenseNet169)

TABLE 3. The overhead of parameters and computation of different comparative models. For each metric, the best value is shown in bold.

Methods	Flops[1]	Params[1]	Methods	Flops	Params
NIMA(IV)[2]	**0.393**	**1.016**	NIMA(D169)[2]	1.168	1.250
NIMA(IRV)[2]	2.665	5.690	D169(SENET)	1.169	1.381
IRV(SENET)	2.674	7.394	D169(C)	1.170	1.381
IRV(C)[2]	2.680	7.394	D169(G)	1.168	1.250
IRV(G)[2]	2.674	5.690	D169(H)	1.168	1.250
IRV(S)[2]	2.674	5.690	D169(S)	1.168	1.250
IRV(H)[2]	2.665	5.690	D169(C&G)	1.170	1.384
IRV(C&G)	2.683	7.395	D169(C&H)	1.170	1.381
IRV(C&H)	2.683	7.394	D169(H&G)	1.168	1.250
IRV(H&G)	2.674	5.690	D169(C&G&H)	1.170	1.384
IRV(C&G&H)	2.683	7.395	MPEMD	2.235	2.409
ASPP FCN-GC	117.6	509.3	SDLA	/	/

[1] The units of Flops and Params are 10^{10} and 10^{7}, respectively.
[2] IV:Inception-V2, IRV:Inception-Resnet-V2, IRV(C):Inception-Resnet-V2(CBAM), IRV(G):Inception-Resnet-V2(SGFM), IRV(H):Inception-Resnet-V2(Hybrid_loss), IRV(S):Inception-Resnet-V2(Softmax cross-entropy), D169:(DenseNet169)

nique for visualizing the important spatial location. Different from the gradients calculated for unordered class, GRAM-CAM tries to take a close look at how network utilizes features for ordered classes task. In Fig.9, we illustrate the visualization results of comparative models. From top to bottom are input image, NIMA(IV), NIMA(D169), NIMA(IRV), IRV(SENET) and IRV(C), it indicates that IRV(C) integrated with CBAM covers more aesthetic factors than the other methods. In comparison with IRV(SENET), it can be clearly seen that IRV(C) learns well to exploit more informative features to represent aesthetics. Note that the visual results of NIMA(SENET) and IRV(C) are also in line with the prediction results reported in Table 1 and Table 2, respectively. Let's take Fig.9(5) for example, the aesthetic score of this image relies on all the distribution of cosplayers. It is clear that only IRV(C) with CBAM could cover all the cosplayers while the others more or less miss ignore the cosplayers who should be considered as a whole. From the observations, we conjecture that CBAM attention module can leverage aesthetic feature to boost the performance.

B. ABLATION STUDY

1) Influence of Baseline Network

In this section, we discuss the influence of baseline network. In NIMA(IRV) and NIMA(D169), we replace the original CNN image extractor (Inception-V2) with Inception-Resnet-V2 and Denset169, respectively, and keep the rest layers unchanged. Clearly from Table 1 and Table 2, NIMA(IRV) outperforms its competitors remarkably on precision metrics. Therefore, two observations can be made:1) The discriminative ability of baseline network plays a core part in IAQA models. 2) The performance of the above models on precision metrics is consistent with the distinguished ability of these three classifier(IV, IRV and D169) in most circumstances.

2) Influence of The Components

In retrospect, our framework consists of three key modules: attention modules, SGFM, and hybrid loss function. These modules are theoretically helpful and beneficial to improve performance. To show their competitiveness, we construct all possible combinations of the above three modules, and then apply them on our two datasets. Now, we respectively analyze the influence of these three components on Alltuu and AVA carefully. At first, we only consider the influence brought by single module and name them as IRV(C), IRV(G), IRV(S), IRV(H), D169(C), D169(S), D169(G) and D169(H). From Table 1, IRV(C) and D169(C) achieve better performance than their competitors, that is to say, regions of interest captured by CBAM can release stronger aesthetic representation power. Compared with NIMA(IRV) and NIMA(D169), IRV(C) and D169(C) gain average 5.2% and 5.6% Precision performance on dataset Alltuu, respectively. In Table 2, we still observe that IRV(C) and D169(C) outperform the others. In comparison with NIMA(IRV) and NIMA(D169), IRV(C) and D169(C) gain average 2.8% and 2% performance on metric Precision, respectively. Now, let's take a closer

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/ACCESS.2020.3015060, IEEE Access

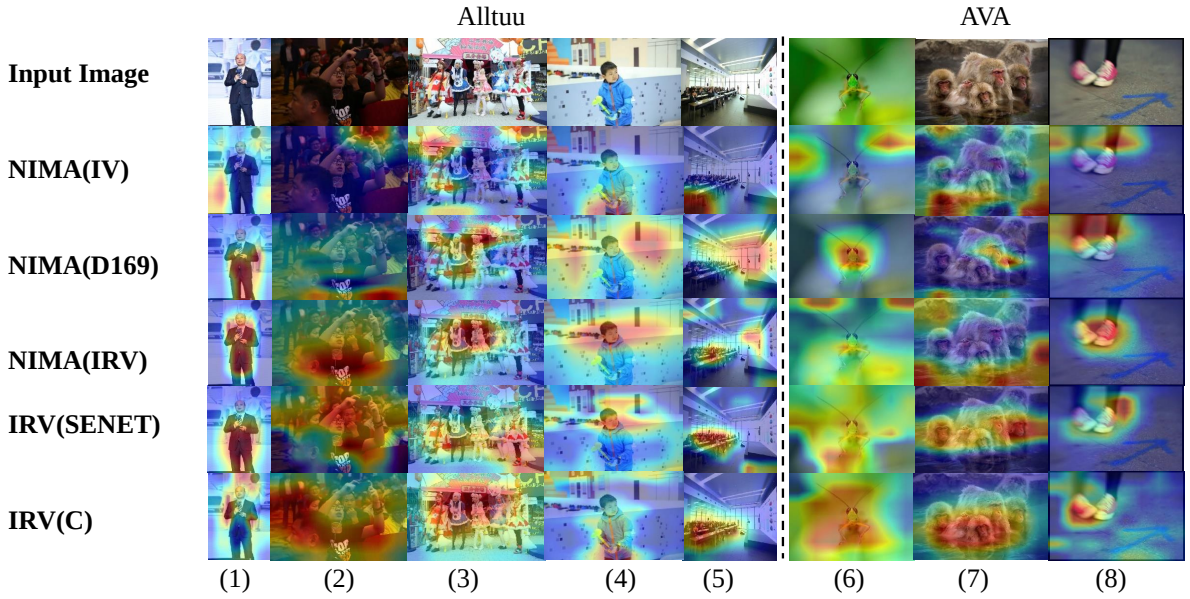

FIGURE 9. The visualization results of comparative methods. Input images are randomly chosen from Alltuu dataset(left columns) and AVA dataset(right columns), respectively.

look at the effectiveness of Hybrid loss. For dataset Alltuu, IRV(H) performs 6.3% better than NIMA(IRV) and D169 performs 11.9% better than NIMA(D169) on metric MAE. If the hybrid loss is replaced with traditional softmax cross-entropy, we find that IRV(S) and D169(S) perform slightly worse than the corresponding IRV(H) and D169(H), respectively. As mentioned in Section IV-C, softmax cross-entropy is better qualified for classification tasks, which is more suitable for the integer requirement of our dataset Alltuu. Whereas, different for classification, our final mean score is computed as $\tilde{V} = \sum_{i=1}^{N} i \times \tilde{\mathbf{S}}_i$ according to algorithm 1. In order to achieve better results, we also resort to EMD loss to force our prediction to be as close as the ground-truth in the case of ordered classes. Herein, one can see that the models equipped with EMD loss (NIMA(IRV) and NIMA(D169)) work worse than the ones equipped with softmax cross-entropy (IR(S) and D169(S)), respectively. For dataset AVA, IRV(H) performs 4.1% better than NIMA(IRV) and D169 performs 3.6% better than NIMA(D169) on metric MAE. In Table 2, different from the phenomena appeared in dataset Alltuu, the models equipped with EMD loss (NIMA(IRV) and NIMA(D169)) show better results than the ones equipped with softmax cross-entropy (IR(S) and D169(S)), respectively. The phenomena might be explained as follows: the scores of sample from dataset AVA are better described in a probability distribution way, which is well fit for the goal off EMD loss. Meanwhile, since the final score is a number rather than a distribution, we add softmax cross-entropy to facilitate the final score closer to some integer label ranging from 1 to 10. Despite that models embedded with SGFM (IRV(G) and D169(G)) and hybrid loss (IRV(H) and D169(H)) fall a little behind the ones with CBAM (IRV(C) and D169(C)), they still make significant progress on precision metrics compared with NIMA(IRV) and NIMA(D169). In particular, we conduct experiments to compare different visual attention modules. As mentioned in Section III-B1, SENET focuses on capturing channel relationship which is fulfilling to locate visual attractive regions. From Table 1 and Table 2, IRV(SENET) and D169(C) are superior to IRV(C) and D169(C) due to the additional spatial attention. Then, we combine two of the above three modules marked as IRV(C&G), IRV(C&H), IRV(H&G), D169(C&G), D169(C&H) and D169(H&G), and then perform them on testing sets. As is seen from Table 1 and Table 2, performance gains significantly compared with the models equipped with only single module. This indicates that the combination of two modules are more descriptive for IAQA. For IRV(H&G) and D169(H&G), their precision performance drop moderately in comparison with the corresponding comparatives, which further prove the validity of CBAM module again. Finally, we test IRV(C&G&H) and D169(C&G&H) equipped with above three modules simultaneously, and find they achieve promising results. This is mainly contributed to relatively excellent CNN-based classifier and the incorporation of three separated modules, which offer strong representation ability of aesthetic evaluation.

3) The Weights in Hybrid Loss

The weights in hybrid loss L play key part in the trade-off between L_s and L_{emd}. To pursue the best weights, we vary α in the range of [0, 1] with the interval of 0.1 and displays the results on datasets Alltuu and AVA in Fig.10. When we only consider L_s or L_{emd}, the MAE increase is in comparison with the best MAE achieved when $\alpha = 0.5$. As α increases, the MAE value increases because it will reduce

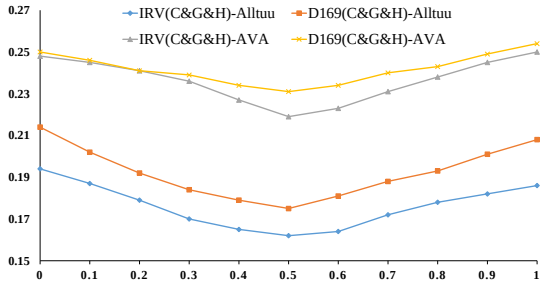

FIGURE 10. Performance variants on metric MAE of ours methods (IRV(C&G&H) and D169(C&G&H)) with different trade-off parameter α and β on the testing sets of datasets Alltuu and AVA.

our models ability. Generally, the trends of the other metrics (MAE, PLCC, SROCC, ACC) are often in consistent with MSE trend in most cases. Thus, in our work, we set $\alpha = 0.5$ to achieve better results through our all experiments.

VI. CONCLUSIONS AND FUTURE WORK.

IAQA is an import application in image modeling and multimedia. In this work, we propose a general framework integrated with attention module, SGFM, and hybrid loss. With the help of these three components, we are able to reflect the attention regions, extract the aesthetic global context information, and optimize IAQA model accurately. The experimental results on datasets Alltuu and AVA demonstrate that the proposed framework are more powerful than previous works in terms of almost all metrics.

In the future, we will investigate a more general and comprehensively IAQA approach, and put emphasis on the following points: 1) Few-shot learning. Thought we solve the over-fitting problem caused by insufficient data with pre-trained models, the burden of collecting large-scale supervised data for industrial needs is still challenging. Thus, we turn our attention to few-shot learning method to achieve better classification. 2) Contaminated labels. Different from the benchmark that are refined and maintained by professionals, numerical samples in real-world application is vulnerable to be contaminated and damaged. Thus, we should learn how to evaluate image quality with only incomplete and contaminated labels. 3) Adaptability ability. A deep neural network can be considered as excellent and adaptive provided that it can be transfered to another field without major modification. Later, we will verify our proposed neural network in the field of fault diagnosis, cropping detection [42] and so on.

ACKNOWLEDGMENT

This research was funded by part by the Young Scientists Fund of the National Natural Science Foundation of China under Grant 61802300, China Postdoctoral Science Foundation Funded Project under grant 2018m643666, Xi'an jiaotong university basic research foundation for Young Teachers under grant xjh012019043, and National Science and Technology Major Project under grant 2019YFB2102501 and 2019YFB2103005.

REFERENCES

[1] Z. Liu, Z. Wang, Y. Yao, L. Zhang, and L. Shao, "Deep active learning with contaminated tags for image aesthetics assessment," IEEE Transactions on Image Processing, 2018.

[2] S. Bianco, L. Celona, and R. Schettini, "Aesthetics assessment of images containing faces," in 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018, pp. 2820–2824.

[3] L. Wang, X. Wang, T. Yamasaki, and K. Aizawa, "Aspect-ratio-preserving multi-patch image aesthetics score prediction," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[4] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 7, pp. 1531–1544, 2018.

[5] D. Liu, R. Puri, N. Kamath, and S. Bhattacharya, "Modeling image composition for visual aesthetic assessment," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[6] N. Ma, A. Volkov, A. Livshits, P. Pietrusinski, H. Hu, and M. Bolin, "An universal image attractiveness ranking framework," in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019, pp. 657–665.

[7] C. Cui, W. Yang, C. Shi, M. Wang, X. Nie, and Y. Yin, "Personalized image quality assessment with social-sensed aesthetic preference," Information Sciences, vol. 512, pp. 780–794, 2020.

[8] D. Liu, R. Puri, N. Kamath, and S. Bhattacharya, "Composition-aware image aesthetics assessment," in The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 3569–3578.

[9] Y. Deng, C. L. Chen, and X. Tang, "Image aesthetic assessment: An experimental survey," IEEE Signal Processing Magazine, vol. 34, no. 4, pp. 80–106, 2017.

[10] K. Yan, X. Tang, and J. Feng, "The design of high-level features for photo quality assessment," in IEEE Computer Society Conference on Computer Vision & Pattern Recognition, 2006.

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

[12] L. K. Wong and K. L. Low, "Saliency-enhanced image aesthetics class prediction," in IEEE International Conference on Image Processing, 2009.

[13] A. Tun? Ozan, S. Aljoscha, and G. Markus, "Automated aesthetic analysis of photographic images," IEEE Transactions on Visualization & Computer Graphics, vol. 21, no. 1, pp. 31–42, 2014.

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

[15] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in Computer Vision & Pattern Recognition, 2011.

[16] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in International Conference on Computer Vision, 2011.

[17] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in Acm International Conference on Multimedia, 2010.

[18] Y. Wu, B. Jiang, and N. Lu, "A descriptor system approach for estimation of incipient faults with application to high-speed railway traction devices," IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017.

[19] Y. Wu, B. Jiang, and Y. Wang, "Incipient winding fault detection and diagnosis for squirrel-cage induction motors equipped on crh trains," ISA transactions, 2019.

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in International Conference on Neural Information Processing Systems, 2012.

[22] W. Wang, M. Zhao, W. Li, J. Huang, C. Cai, and X. Xu, "A multi-scene deep learning model for image aesthetic evaluation," Signal Processing Image Communication, vol. 47, no. C, pp. 511–518, 2016.

[23] K. Apostolidis and V. Mezaris, "Image aesthetics assessment using fully convolutional neural networks," in International Conference on Multimedia Modeling, 2019.

[24] X. Zhang, X. Gao, W. Lu, and L. He, "A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction," IEEE Transactions on Multimedia, vol. 21, no. 11, pp. 2815–2826, 2019.

[25] X. Tian, D. Zhe, K. Yang, and M. Tao, "Query-dependent aesthetic model with deep learning for photo quality assessment," IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 2035–2048, 2015.

[26] L. Xin, L. Zhe, H. Jin, J. Yang, and J. Z. Wang, "Rating pictorial aesthetics using deep learning," IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 457–466, 2015.

[27] S. Ma, J. Liu, and C. Wen Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4535–4544.

[28] L. Xin, L. Zhe, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in IEEE International Conference on Computer Vision, 2016.

[29] H. Roy, T. Yamasaki, and T. Hashimoto, "Predicting image aesthetics using objects in the scene," in Proceedings of the 2018 International Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia. ACM, 2018, pp. 14–19.

[30] H. Talebi and P. Milanfar, "Nima: Neural image assessment," IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 3998–4011, 2018.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[33] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[35] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6306–6314.

[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[37] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[38] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in Advances in neural information processing systems, 2017, pp. 971–980.

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

[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al., "Imagenet large scale visual recognition challenge," International journal of computer vision, vol. 115, no. 3, pp. 211–252, 2015.

[41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[42] A. Wang, Y. Xu, X. Wei, and B. Cui, "Semantic segmentation of crop and weed using an encoder-decoder network and image enhancement method under uncontrolled outdoor illumination," IEEE Access, vol. 8, pp. 81 724–81 734, 2020.