M+MNet: A Mixed-Precision Multibranch Network for Image Aesthetics Assessment

HE Shuai[1], LIU Limin[1], WANG Zhanli[2], LI Jinliang[2], MAO Xiaojun[2], MING Anlong[1]

(1. Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202503011

https://kns.cnki.net/kcms/detail/34.1294.TN.20250905.1148.002.html, published online September 5, 2025

Manuscript received: 2024-11-13

Abstract: We propose Mixed-Precision Multibranch Network (M+MNet) to compensate for the neglect of background information in image aesthetics assessment (IAA) while providing strategies for overcoming the dilemma between training costs and performance. First, two exponentially weighted pooling methods are used to selectively boost the extraction of background and salient information during downsampling. Second, we propose Corner Grid, an unsupervised data augmentation method that leverages the diffusive characteristics of convolution to force the network to seek more relevant background information. Third, we perform mixed-precision training by switching the precision format, thus significantly reducing the time and memory consumption of data representation and transmission. Most of our methods specifically designed for IAA tasks have demonstrated generalizability to other IAA works. For performance verification, we develop a large-scale benchmark (the most comprehensive thus far) by comparing 17 methods with M+MNet on two representative datasets: the Aesthetic Visual Analysis (AVA) dataset and FLICKR-Aesthetic Evaluation Subset (FLICKR-AES). M+MNet achieves state-of-the-art performance on all tasks.

Keywords: deep learning; image aesthetics assessment; multibranch network

Citation (Format 1): HE S, LIU L M, WANG Z L, et al. M+MNet: a mixed-precision multibranch network for image aesthetics assessment [J]. ZTE Communications, 2025, 23(3): 96 – 110. DOI: 10.12142/ZTECOM.202503011

Citation (Format 2): S. He, L. M. Liu, Z. L. Wang, et al., "M+MNet: a mixed-precision multibranch network for image aesthetics assessment," *ZTE Communications*, vol. 23, no. 3, pp. 96 – 110, Sept. 2025. doi: 10.12142/ZTECOM.202503011.

1 Introduction

Assessing image aesthetics is challenging because it requires correctly defining the aesthetic features in an image while precisely evaluating the subjective aesthetics. For example, a classification model can easily identify the tree in Fig. 1a, but current aesthetic assessment models may have difficulty describing why the aesthetics of the tree would earn this image more than 20 000 views and 1 000 likes on the photo-sharing website Flickr. Researchers[1 – 4] have demonstrated that the background, composition, and visual weight balance of images are key factors for its beauty. Therefore, background information is crucial for image aesthetics assessment (IAA) tasks and should be considered in related network designs.

However, few existing convolutional neural network (CNN)-based network designs address this issue. As shown in Fig. 1a, current network layers are designed to focus on regions of high activations in the feature map, and commonly employ pooling methods to discard low activations during downsampling, po-

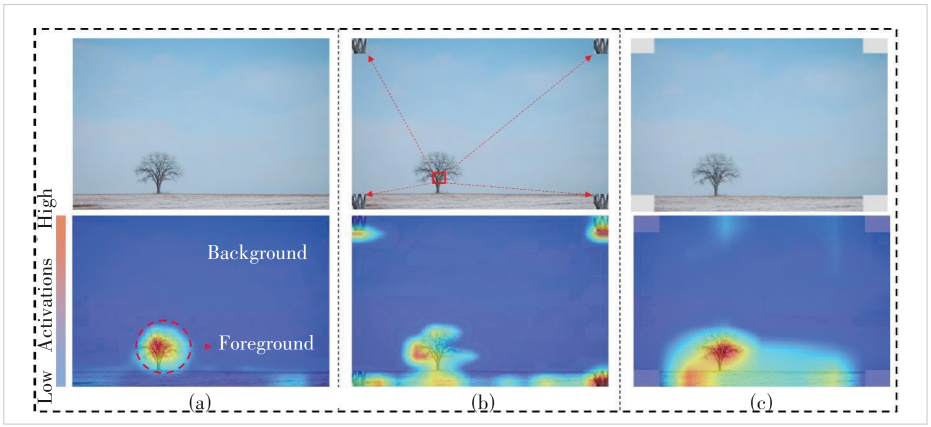


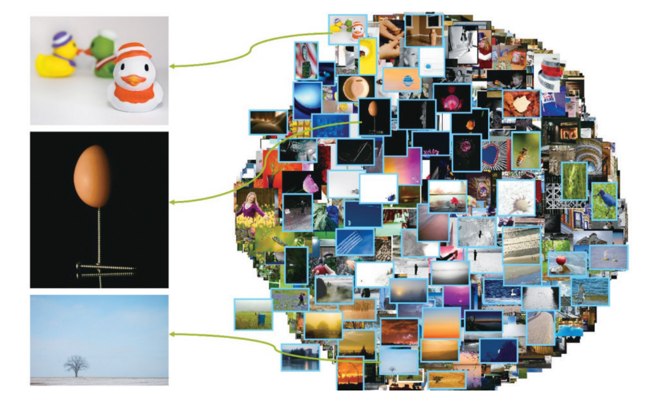
Figure 1. Visualizations of feature map activations generated via Grad-CAM[5]. Our model was pretrained on ImageNet[6] to initialize the weights: (a) Background and foreground information in the image correspond to low and high activations in the original feature map; (b) by copying a small part of the salient region to each of the four corners, the attention area is enlarged; (c) the proposed data augmentation method Corner Grid can be used to markedly increase the attention area

This work is supported by the National Natural Science Foundation of China under Grant No. 62502040 and the ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20230700001.

tentially losing important background information. Notably, existing models exhibit large prediction errors for certain images, such as images with small proportions of salient objects relative to a large background or images whose aesthetics are closely related to their backgrounds, which we call "background-sensitive" samples. For example (Fig. 2), in Neural Image Assessment (NIMA)[7] trained on the large-scale Aesthetic Visual Analysis (AVA) dataset[8], approximately 5.2% of the training samples and approximately 15% of the test samples are background-sensitive, degrading model performance.

To solve the above problem, we have made the following efforts: 1) We exploit a simple Multibranch Network (MNet) with two dedicated pooling methods. These pooling methods normalize all feature map activations to obtain two prior weights. From a human perspective, such weights tend to focus on background information or foreground information; from a model perspective, these weights are used to aggregate the low or high activations. Thus, the mechanism of the proposed pooling methods conforms to the common sense that the background information and foreground information correspond to low and high activations in the feature map[9], respectively. Specifically, one of the weights that tends to preserve low activations is assigned to obtain background information. 2) We introduce an unsupervised data augmentation method named Corner Grid to seek more relevant background information; the motivation is illustrated in Fig. 1. Previous works indicate that in classical computer vision tasks, by changing part of the information in an image (Fig. 1b), a CNN can effectively learn the information that was originally less sensitive, thereby increasing the attention area[10 – 12]. Through the convolution operations of CNN-based methods, the focus can be spread from neighboring pixels to cover more areas. Based on this characteristic, we propose a data augmentation method suitable for IAA tasks, which works by changing the pixel values at the four corners of an image to increase the attention area (Fig. 1c). Similar to HE et al.'s masked autoencoder (MAE)[13], Corner Grid is essentially a mask, and it encourages the model to learn useful features from the background and understand beyond image background statistics.

In addition to the limitations of network design, IAA models are often compromised by the constraints of the existing training strategies. Most existing IAA models have been pretrained on the ImageNet dataset[6] to initialize their weights, meaning that the size of the image inputs used for pretraining is 224×224. To prevent misalignment of the weights transferred to aesthetic tasks, these methods continue to use this input size by default. However, this size is not the optimal size for IAA tasks, and its use can lead to incomplete extraction of aesthetic information and impair the performance of IAA models. Although using higher-resolution inputs can preserve more of the available aesthetic information, this will lead to high memory consumption while limiting the training speed. Moreover, because rater subjectivity can generate noise in the ground-truth labels of an aesthetic benchmark, solving the IAA problem typically requires learning from a noisy raw score distribution (Fig. 3); consequently, relatively long training times are already needed to achieve better generalization ability. Therefore, most previous works on IAA have faced difficulties in balancing performance and training costs.



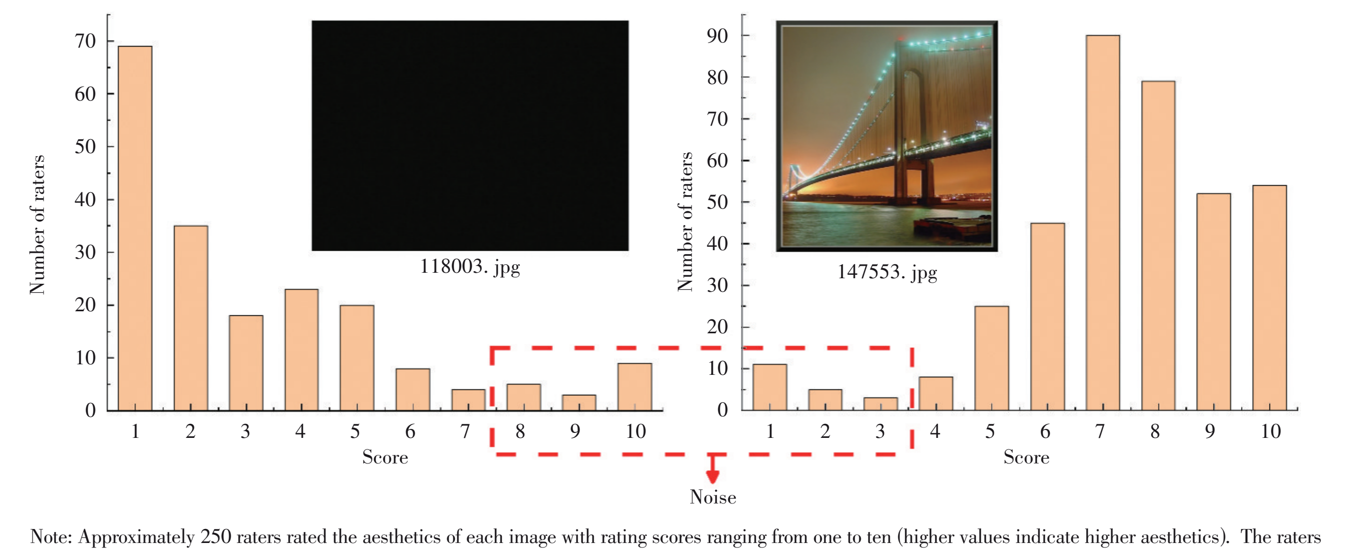
Note: The aesthetics of these images are usually closely related to the background. We call such images "background-sensitive" samples.

Figure 2. Visualization of images in the AVA dataset with a large absolute error between the ground truth and the predicted score (absolute error ⩾ 1)

High-resolution input can compensate for the aesthetic detail lost through the use of low-resolution input[14]. However, high-resolution demands increase memory consumption and reduce training speed due to data transmission, storage, and arithmetic needs[15]. Motivated by these considerations, our training strategy uses multiple training stages to achieve a transition between low- and high-resolution input and leverages a mixed-precision approach to reduce the memory consumed for data representation. The training system is designed based on this training strategy from the bottom up and thus can fundamentally alleviate the abovementioned dilemma. To further achieve high performance during training, we adopt three techniques to alleviate performance degradation caused by the mixed-precision approach and improve the traditional Earth mover's distance (EMD) loss by rebalancing the loss contributions based on the notion of ground-truth consistency.

The main contributions of this work are as follows:

- To effectively extract aesthetic information from images, especially background information, we design a novel multibranch network equipped with two dedicated pooling methods. In addition, an unsupervised data augmentation method is proposed to seek more relevant background information.
- To address the dilemma between performance and training costs, an improved mixed-precision training strategy and an improved loss function are presented. The proposed method is ten times faster than previous methods and reduces GPU memory usage by approximately 19.37% while achieving state-of-the-art (SOTA) performance.
- To provide a comprehensive evaluation for the commu-



Note: Approximately 250 raters rated the aesthetics of each image with rating scores ranging from one to ten (higher values indicate higher aesthetics). The raters gave some high scores to "118003. jpg" for a pure black image and some low scores to "147553. jpg" for a good one; these apparently incorrect scores are noise.

Figure 3. Samples selected from the AVA dataset, along with plots of their ground-truth score distributions

nity, we compare 17 SOTA baselines on two representative datasets, AVA[8] and FLICKR-Aesthetic Evaluation Subset (FLICKR-AES)[16], making this work the most complete IAA benchmark to date.

- Our proposed techniques, such as the pooling methods, the unsupervised data augmentation method, and the training strategy, can independently be embedded in existing methods or training processes to solve possible stumbling blocks on IAA tasks.

2 Related Work

2.1 Image Aesthetics Assessment

General IAA involves three tasks: binary classification, aesthetic score regression, and score distribution prediction. Due to the complexity of manual feature extraction and the lack of training data, early methods[17 – 18] treated IAA as a binary classification task (aesthetically positive or negative). Recently, CNN methods[7, 19 – 21] have been proposed for binary aesthetic classification. These efforts are based on extracting multilevel aesthetic features and using the standard cross-entropy (CE) loss to train an IAA model. Benefiting from the large-scale AVA dataset[8], researchers have also been able to obtain reasonable performance on the more challenging aesthetic score regression task[21 – 24]. In addition, KONG et al.[24] and TALEBI et al.[7] reported their results on AVA in terms of the Spearman rank correlation coefficient (SRCC) metric, which is a natural way to evaluate the ranking loss.

Although such methods have achieved great success, recent evidence reveals that directly predicting aesthetic scores (score regression) obscures the diversity of human opinions[7, 25]. For example, each image in the AVA dataset[8] was rated by an average of 250 raters, but the average aesthetic score does not reflect the subjective preference of all individual raters. Some researchers have noted this limitation and proposed using the EMD loss[26 – 29] for the score distribution task, and this approach shows promising performance[7, 29 – 32]. To further model subjective preferences, some works[16, 32 – 34] have proposed a personality-assisted multitask framework for personalized aesthetic tasks. However, overemphasizing an individual's subjective preference degrades SRCC performance in both general and personalized aesthetic tasks. The main reason for this shortcoming is that some raters will give an apparently incorrect score that is too high or too low (Fig. 3), and it is difficult and time-consuming for IAA models to fit such minority (or noisy) opinions.

To alleviate the problem mentioned above, it is preferable for IAA models to only focus on majority opinions; thus, we present the rebalanced EMD (Re-EMD) loss function to reweight the loss distribution to help the network focus more on majority opinions during training.

2.2 Multibranch Networks

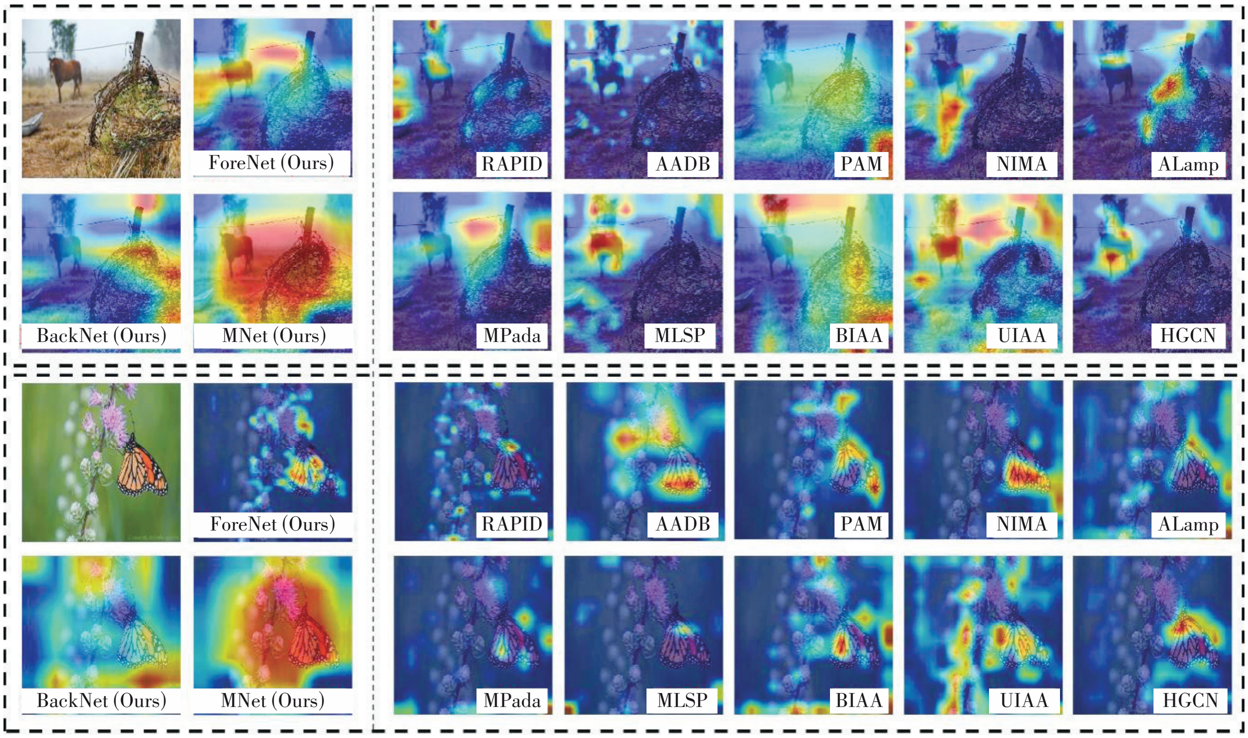
Despite the lack of firm rules governing aesthetic appeal, certain aesthetic features are believed by many to be more pleasing to humans than certain other features. Multibranch networks are popular methods for the extraction of aesthetic features at different levels. Both local and global features were regarded as crucial aesthetic information in early multibranch-based methods[19 – 20]. Similar to these works, MA et al.[21] adopted attribute graphs to represent structured groups with local and global layouts, and ZHANG et al.[35] focused on both the global composition and local fine-grained details. In other studies[25, 36], researchers have reported that visual and textural features are the key features of interest in IAA tasks. Notably, the existing multibranch networks can easily focus on salient objects or semantically meaningful content but respond only slightly to background regions without significant features; however, unlike the tasks of image classification and object recognition, which often focus on salient objects, IAA is also heavily dependent on background information[37 – 38]. Nevertheless, as shown in Fig. 4, typical IAA models focus on salient objects but disregard the background, which may either enhance or weaken the aesthetics of the image, thereby limiting the performance of these methods on IAA tasks.

To solve the above issue, we design a simple multibranch network called MNet, which is equipped with two dedicated pooling methods for extracting salient and background information. Furthermore, we explore an unsupervised data augmentation method called Corner Grid to increase the model's attention to background information. Experimental results show that the proposed method achieves better performance than previous methods.

2.3 Reduced Precision Training

For a given network structure, the total training costs (e.g., memory consumption and training time) depend on the input resolution, batch size, and precision utilized by the system. Using low-resolution images as the input results in a loss of fine-grained details, while using a small batch size causes poor model generalization. In recent studies, reduced precision representations have been applied to reduce the training costs.

COURBARIAUX et al.[40] converted the weights to a binary format but maintained the gradients and activations as single-precision values during the training process. HUBARA et al.[41] reduced both weights and activations to low-precision values (<6 bit) for CNN training. HE et al. [42] applied the same method for recurrent neural network training. ZHOU et al.[43] further used low-precision representations of the weights, activities, and gradients. However, all of these approaches lead to performance degradation when applied to large models or datasets. Since a low-precision format has a narrower dynamic range than a high-precision format, a key issue is how to avoid representation errors, such as overflow, underflow, and rounding errors. When a value in the single-precision (FP32) format is converted to the half-precision (FP16) format, overflow will occur if the number is greater than 65 504, and underflow will occur if the number is less than 6×10^{-8}. The FP16 format also has a narrower dynamic range than FP32, which may cause



Note: Our MNet method can effectively improve the attention to background areas related to salient objects, thus yielding results that are more consistent with human perception.

AADB: Aesthetics and Attributes Database
ALamp: Adaptive Layout-Aware Multi-Patch Deep Convolutional Neural Network
BIAA: Bilevel Gradient Optimization Image Aesthetics Assessment
HGCN: Hierarchical Layout-Aware Graph Convolutional Network
MLSP: Multi-Level Spatially Pooled Features
MNet: multibranch network
MPada: Attention-Based Multi-Patch Aggregation
NIMA: Neural Image Assessment
PAM: Personalized Aesthetics Model
RAPID: Rating Pictorial Aesthetics Using Deep Learning
UIAA: Unified Image Aesthetic Assessment

Figure 4. Activation maps comparing benchmark IAA models (Table 1) and our proposed method through fused 2D feature maps of the last layers of these models

rounding errors during weight updates. For example, $2^{-24}+2^{-36} \approx 2^{-24}$, and any value with a magnitude smaller than 2^{-24} becomes zero in FP16. To solve these problems, MICIKEVICIUS et al.[15] and PURI et al.[39] proposed a mixed-precision training strategy to quickly train large-scale models. The core of the existing methods could be summarized as a "skipping" strategy. This kind of strategy attempts to store and transport data in the FP16 format, and if overflow or underflow occurs on a certain data batch, it will skip (discard) the current data batch and attempt to represent the data in the next data batch using a higher-precision format.

However, applying a mixed-precision approach to train IAA models is especially challenging, because the aesthetic appeal of an image is a subjective property while outlier opinions may appear and then the quality of the ground truth is consequently not high (Fig. 3). This situation causes instability during initial training, and IAA models usually require a long time to reduce the loss to a meaningfully smaller value. Therefore, the gradients often exceed the range that can be represented in the FP16 format, resulting in an excessive number of ineffective data batches, as shown in Fig. 5a. To achieve a balance between performance and training speed, we adopt three techniques to mitigate the problems caused by mixed-precision training: gradient monitoring, automatic loss scaling, and accumulation in FP32. Gradient monitoring is performed as a precaution to enable the network to enter mixed-precision training in a more stable state (Fig. 5b), while the other two techniques are applied to correct the representation errors that arise in mixed-precision training.

3 Methods

3.1 Design of Multibranch Network

Based on the characteristics of IAA tasks, our network architecture is designed as shown in Fig. 6. We first introduce pooling methods for extracting foreground and background information, along with strategies to fix the output size of these pooling methods regardless of input size variations. Second,

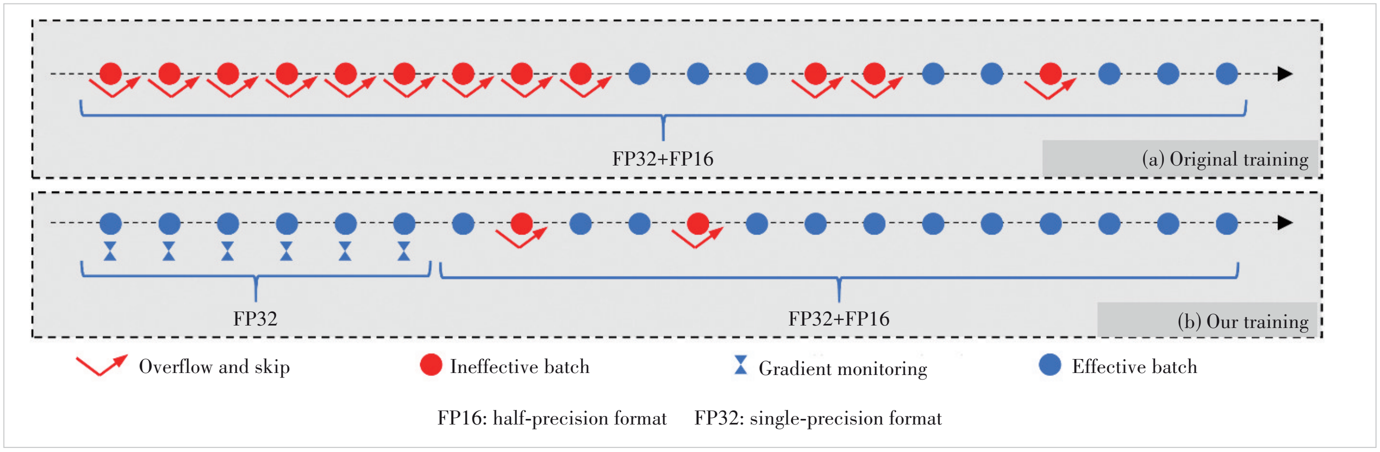
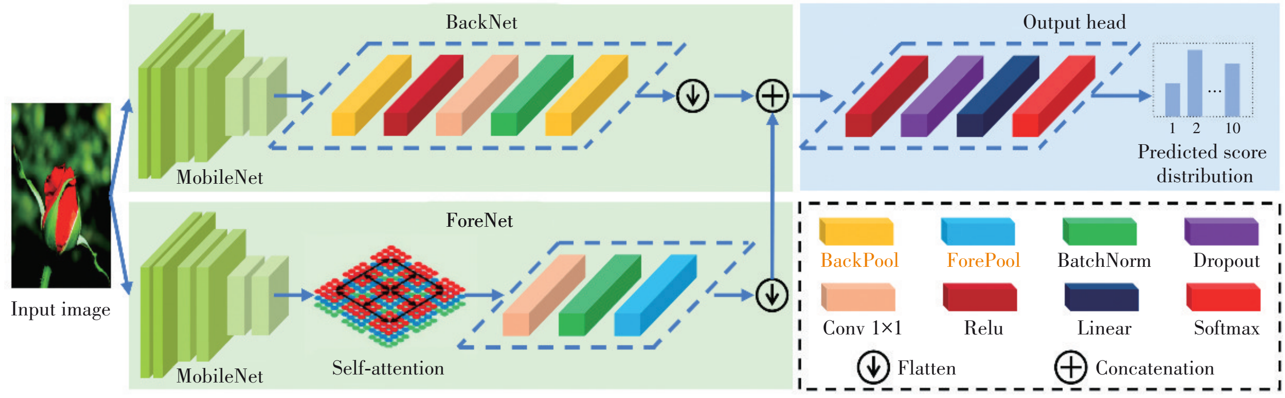


Figure 5. Comparison of conventional mixed-precision training[15, 39] and our training method: Overflowed batches are skipped until stable gradients trigger phase transition



Note: The salient objects and background information are extracted by ForePool and BackPool in ForeNet and BackNet, respectively. After flattening, the features are sent to the output head to predict the score distribution.

Figure 6. Overall architecture of the proposed MNet

we use self-attention mechanisms to enhance the network's understanding of the relationships among multiple subjects in the foreground. Finally, a 1×1 convolution kernel is adopted to balance the information output by the multibranch network.

3.1.1 Design of Pooling Methods

We design MNet with two sub-branches, ForeNet and BackNet, both adopting partial layer structures from MobileNetV2[44] for feature extraction. However, neither sub-branch contains the original pooling layers, as the original max pooling and average pooling layers prove ineffective for preserving background information. Accordingly, we develop two dedicated pooling methods.

For an input image X, we extract the feature maps before any pooling layers. Each value in a feature map represents an activation x_i, and we assign a weight t_i to each activation. Then, the extracted feature maps are passed through pooling layers, in which the output for each pooling kernel region Ω is calculated as $\sum_{i \in \Omega} t_i \cdot x_i$. Since salient information usually corresponds to relatively large activation values in feature maps, it can be inferred that background information corresponds to relatively small activation values[8]. Thus, for ForePool, the output weight of each x_i is defined as follows:

$$t_i = \frac{e\left(x_i\right)}{\sum_{j \in \Omega} e\left(x_j\right)} \tag{1},$$

where the exponential function $e(\ )$ is used to enlarge the activation values to better distinguish background and salient information. This pooling method ensures that the higher activations corresponding to salient objects will play a dominant role while still preserving some background information. In contrast, for BackPool, the output weights are calculated as follows:

$$t_i = 1 - \frac{e\left(x_i\right)}{\sum_{j \in \Omega} e\left(x_j\right)} \tag{2}.$$

In this way, the background information associated with lower activations is extracted while still ensuring that some salient information is retained. Compared with classical average or max pooling (Fig. 7), our pooling methods are more balanced in extracting important information and secondary information, depending on the tasks of different sub-branches.

To enhance the robustness of our MNet to different input sizes, we design ForePool and BackPool to adaptively pool the arbitrarily sized input $X^{c_{\text{in}} \times h_{\text{in}} \times w_{\text{in}}}$ to a desired feature map size $D^{c_{\text{out}} \times h_{\text{out}} \times w_{\text{out}}}$, where c_{in} and c_{out} denote the numbers of input and output channels, respectively, and $h_{\text{in}} \times w_{\text{in}}$ and $h_{\text{out}} \times w_{\text{out}}$ represent the input and output feature map sizes, respectively. Based on the desired feature map size $D^{c_{\text{out}} \times h_{\text{out}} \times w_{\text{out}}}$ and the input resolution $h_{\text{in}} \times w_{\text{in}}$, our pooling methods dynamically adjust the strides $\left(s_h, s_w\right) = \left(\left\lfloor \frac{h_{\text{in}}}{h_{\text{out}}} \right\rfloor, \left\lfloor \frac{w_{\text{in}}}{w_{\text{out}}} \right\rfloor\right)$ and the adaptive kernel dimensions $\left(k_h, k_w\right) = \left(\left(h_{\text{in}} - \left(h_{\text{out}} - 1\right) \times s_h\right), \left(w_{\text{in}} - \left(w_{\text{out}} - 1\right) \times s_w\right)\right)$. This ensures a fixed output size during training, and specifically, the padding size is set to 0.

3.1.2 Understanding Relationships Among Subjects

According to previous works[28], understanding the relationships among multiple subjects in an image is important for IAA tasks since an appropriate arrangement of visual ele-

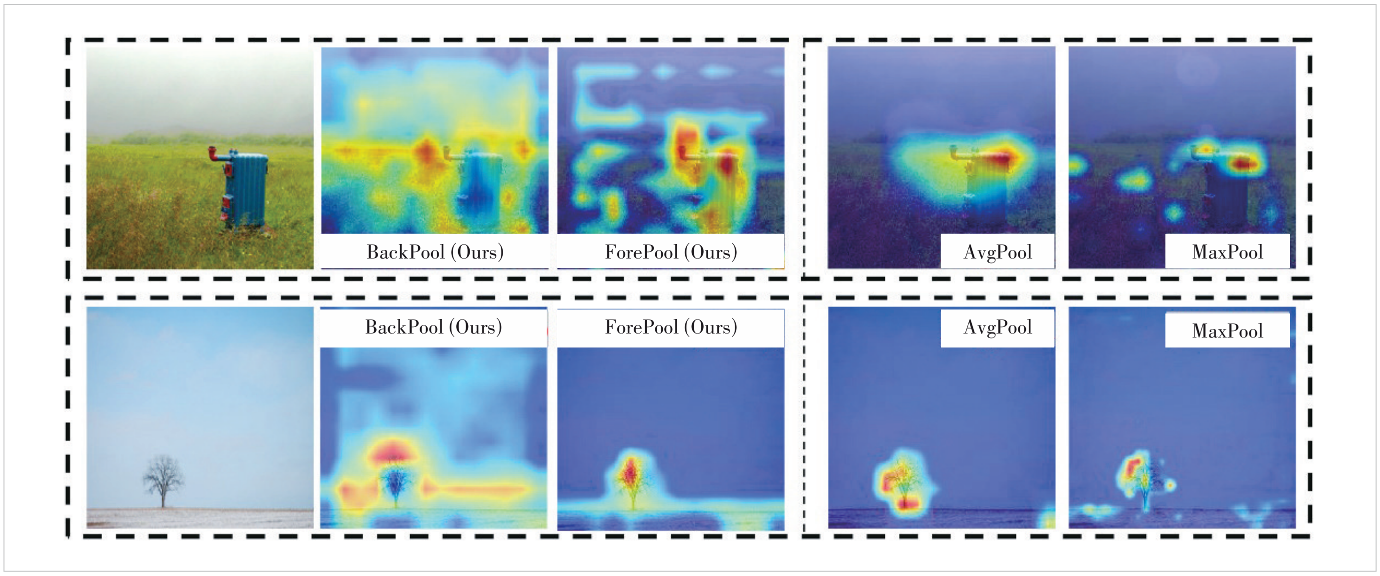


Figure 7. Different activation maps are obtained when the pooling methods in NIMA are replaced with the proposed BackPool and ForePool methods, or with the traditional average and max pooling methods

ments in an image can benefit visual balance and harmony.

Instead of utilizing the power of complex networks, we add a self-attention layer[45] to ForeNet to gain an understanding of these relationships. A self-attention mechanism can detect the relationships among key foreground regions (each usually containing some salient objects or semantically meaningful content) while enabling the model to pay different levels of attention to different objects[45–46]. Through the power of self-attention, the salient objects extracted by the backbone network can be carefully coordinated with fine details in distant portions of the image. Considering that the original multiple-subject relationships of the image may be incomplete after the downsampling process, we place the self-attention layer before the ForePool layer.

3.1.3 Balancing Extracted Information

Another problem that needs to be solved in our MNet is how to aggregate the feature maps extracted in different sub-branches. In previous works on IAA[14, 23], the feature maps with different channels have simply been concatenated. However, the channels with different numbers mean that the information contributed by each sub-branch may not be balanced, which may cause the importance of sub-branches with rich channels to be over-weighted and complicate the training process[47].

To balance this spatial information, we add a commonly utilized 1×1 convolution kernel[48] to MNet. As a cross-channel pooling structure, this kernel enables cross-channel spatial information interaction and cascaded cross-channel parametric pooling of the features extracted by the two sub-branches in a normal convolution layer. Thus, it can be ensured that the aesthetic information from two sub-branches is aggregated after the reduction and equalization of the number of channels.

3.2 Loss Function

Generally, the ground truth in an IAA dataset consists of the score distribution (Fig. 3), and our network aims to predict this distribution. Since the EMD loss penalizes misclassifications based on class distances, it is well-suited for measuring the distances between ground-truth and predicted distributions, as demonstrated in previous works[7]. Given a ground-truth distribution $p = \left(p_1,\cdots,p_N\right)$ and a predicted distribution $\hat{p} = \left(\hat{p}_1,\cdots,\hat{p}_N\right)$, with N ordered classes, the original EMD loss can be expressed as follows:

$$\mathrm{EMD} = \left(\frac{1}{N}\sum_{k=1}^{N}\left|f_p(k) - f_{\hat{p}}(k)\right|^{\gamma}\right)^{\frac{1}{\gamma}} \tag{3},$$

where $f_p(k)$ is the cumulative distribution function, calculated as $\sum_{i=1}^{k} p_i$, and γ is used to penalize the Euclidean distance and has usually been set to 1 or 2 in previous works. However, due to the strong subjectivity of IAA, there is typically some noise in the ground-truth distribution caused by the minority opinions of a few raters (Fig. 3), and it is difficult for the model to fit these opinions. A key issue is how to weaken the contribution of the noise to the loss. To solve this problem, our main improvement to the EMD loss is to introduce the notion of ground-truth consistency by multiplying by a normalization weight related to the ground-truth distribution. Thus, we refer to this loss function as the rebalanced EMD (Re-EMD), which is formulated as follows:

$$\text{Re-EMD} = \left(\frac{1}{N}\sum_{k=1}^{N}\left|\boldsymbol{M}(p)\cdot\left(f_p(k) - f_{\hat{p}}(k)\right)\right|^{\gamma}\right)^{\frac{1}{\gamma}} \tag{4},$$

where the weight $\boldsymbol{M}(p) = \left(\frac{\left(p_1,\cdots,p_N\right)}{\sum_{j=1}^{N} p_j} + \beta\right)\cdot\alpha$, with β being a small constant preventing a weight of zero. Because the weight after rebalancing ranges between 0 and 1, we amplify it by α. The design of our loss function is based on the following consideration: in the ground-truth distribution, the more raters vote for a certain score, the more likely it is that this score represents the image's true rating. Thus, we make the network give priority to the opinion label given by the majority of raters and pay less attention to unusual labels, thereby enhancing the consistency of the loss contribution of the ground truth.

3.3 Mixed-Precision Training

The gradients during early IAA model training often exceed the range that FP16 can represent (Fig. 5b); therefore, it is not practical to apply mixed-precision training from the beginning. A simple and effective way is to appropriately delay the time of entry for mixed-precision training until the gradients can be represented in FP16 most of the time. To allow the system to automatically decide when to enter mixed-precision training, we define a threshold value θ:

$$\theta = \lambda_1 O^2 - \lambda_2 E \tag{5},$$

where E is the total number of training epochs and O represents the number of epochs among the five most recent epochs in which gradient overflow has occurred, which can be automatically calculated during training; λ_1 and λ_2 are predefined hyperparameters that control the degree of restriction. We monitor the gradients during training. If $\theta \le 0$ is detected, meaning that gradient overflow occurs sufficiently infrequently and the model is considered relatively stable, the system can switch to the mixed-precision training format in the next epoch, as shown in Fig. 5b. Gradient monitoring is performed as a precaution to avoid entering the mixed-precision training stage when the model is not yet sufficiently stable. As the number of training epochs increases, the network will eventually enter the mixed-precision training stage despite minor representation errors.

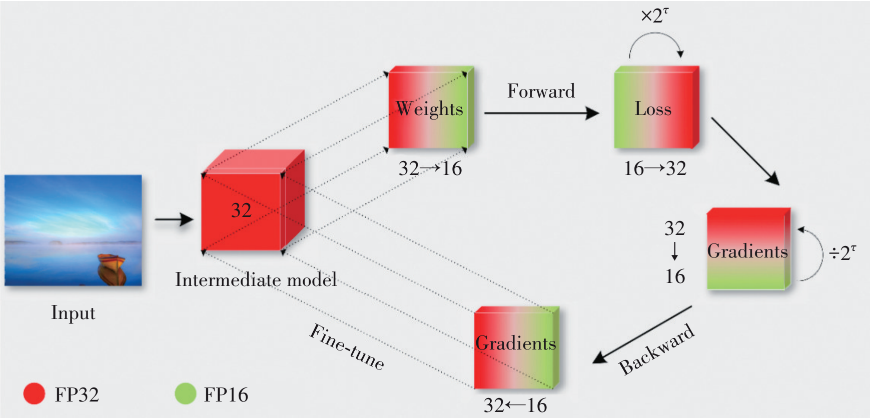
Two techniques are applied to correct representation errors arising in mixed-precision training: automatic loss scaling and FP32 accumulation, as illustrated in Fig. 8. Before mixed-precision training, we convert the intermediate weights to FP16 while maintaining an FP32 master copy. The FP16 weights are then used throughout the entire forward process, but the loss is calculated in FP32. To prevent small gradients from vanishing during backpropagation, we scale the loss by a factor of 2^{τ} ($\tau \leqslant 20$), following previous works[15, 39]. By the chain rule of backpropagation, the intermediate gradients are automatically scaled by 2^{τ}, mitigating rounding errors. Before backpropagation, we divide the final gradients by 2^{τ} and convert them to FP16. However, when an overflow occurs, we abandon the current batch and reduce τ in the next batch; otherwise, backpropagation proceeds normally. To avoid rounding errors during weight updates, gradients are converted to FP32 and accumulated into the FP32 master weights.

In summary, we use the FP16 format to perform most operations in order to reduce memory consumption and boost the training speed, then we use FP32 for operations that would otherwise cause a decrease in accuracy. Thus, our Mixed-Precision MNet (M+MNet) can be trained more quickly.

3.4 Corner Grid

CNN-based methods prioritize regions that represent foreground information, possess unique features (e. g., lines, curves), and contain different pixels[49]. Based on this characteristic, we augment background pixels to encourage our model to learn useful features from the background. However, one prerequisite is that these pixel changes preserve the subject's visual coherence (Fig. 1b). To achieve this goal, we propose Corner Grid, an unsupervised data augmentation method that extracts the average pixel values in the whole image and then overwrites certain grid cells with these pixel values. These average pixels contain salient foreground information, diverting the model's attention to spread toward these grid cells.

We express the size of one grid cell as $(w_g, h_g) = (w_{\text{in}} r, h_{\text{in}} r)$, where w_{in} and h_{in} are the width and height of the input, respectively, and r is the scale of the mask grid with respect to the input (which is the same in both the horizontal and vertical directions). A grid cell can be defined using its top-left and bottom-right pixel positions. If the coordinates of the top-left corner of the image are (0, 0), the coordinate positions of the four grid cells can be given as follows: $(0, 0, w_g, h_g)$, $(w_{\text{in}} - w_g, 0, w_{\text{in}}, h_g)$, $(0, h_{\text{in}} - h_g, w_g, h_{\text{in}})$, and $(w_{\text{in}} - w_g, h_{\text{in}} - h_g, w_{\text{in}}, h_{\text{in}})$.



Note: Weights, activations, and gradients are stored in FP16; data are converted to FP16 during forward and backward passes and for multiplication operations, while FP32 is used to copy master weights for updates and accumulation.

FP16: half-precision format FP32: single-precision format

Figure 8. Mixed-precision training process

We use the Gray World (GW)[50] algorithm to compute and assign pixel values for these grid cells. The GW algorithm is based on the assumption that the color in each sensor channel averages out to gray over the entire image. This algorithm can adjust the pixel values based on the pixel distribution of the whole image. Thereby, the filled pixels will not visually conflict too much with the color of the main body of the image while implicitly preserving foreground information. Let $I_r(x, y)$, $I_g(x, y)$, and $I_b(x, y)$ denote the red, green, and blue channels, respectively, where x and y denote the pixel position indices. The average pixel value of the whole image in these three channels can be calculated as $W = (\bar{R} + \bar{G} + \bar{B})/3$, where

$$(\bar{R}, \bar{G}, \bar{B}) = \frac{1}{wh} \sum_{x=1}^{w} \sum_{y=1}^{h} \left(I_r(x, y), I_g(x, y), I_b(x, y) \right) \tag{6}.$$

We then adjust the red, green, and blue channels' pixel values of each corner grid cell as follows:

$$\hat{I}_r(x, y) = \frac{W}{\bar{R}} \cdot I_r(x, y) \tag{7},$$

$$\hat{I}_g(x, y) = \frac{W}{\bar{G}} \cdot I_g(x, y) \tag{8},$$

$$\hat{I}_b(x, y) = \frac{W}{\bar{B}} \cdot I_b(x, y) \tag{9}.$$

The proposed Corner Grid method can be easily implemented in PyTorch or TensorFlow, and we provide an example implementation in our code.

4 Experimental Results

4.1 Settings

4.1.1 Benchmark Datasets

We evaluated models on two representative datasets, AVA[8] and FLICKR-AES[16], which are the largest general and personalized aesthetic datasets for IAA tasks, respectively. The AVA dataset contains approximately 250 000 images, and each image is associated with a distribution of scores in a range of 1 – 10 rated by approximately 250 raters. The FLICKR-AES dataset consists of 40 000 images whose aesthetic scores range from 1 to 5 to reflect different levels of image aesthetics, and each image was rated by 5 raters. For the AVA dataset, we split the images into training (80%) and test (20%) datasets, as in previous general IAA works[7, 14, 21, 29, 31, 51]. For the FLICKR-AES dataset, we used the same training and test datasets used in previous works on personalized IAA[16, 32 – 34].

4.1.2 Benchmark Models

In accordance with two criteria, recency of publication and representativeness of the pipeline, we selected 17 SOTA models[7, 14, 16, 19 – 21, 23 – 24, 27 – 29, 32 – 34, 51 – 53] for evaluation on the AVA dataset. In addition, we selected four specialized designs[16, 32, 33 – 34] oriented toward personalized aesthetics assessment for performance evaluation on the FLICKR-AES dataset.

4.1.3 Evaluation Metrics

We adopt three popular evaluation metrics: SRCC[7], Linear Correlation Coefficient (LCC)[7], and binary classification accuracy (Acc). For Acc, images with average scores less than or equal to five are deemed aesthetically negative. AVA evaluation additionally includes EMD loss[7]. Although most previous IAA methods trained on the AVA dataset have shown improvements in binary classification accuracy, there are some problems with this metric. In particular, disparate predicted scores for the same image may all be considered correct predictions; for example, a predicted score of either 5.1 or 8.1 is considered correct for an image with a positive aesthetic assessment. As HOSU et al. [14] demonstrated, higher SRCC/Acc ratios generalize better across the entire score range. Therefore, the SRCC/accuracy ratio was reported on our benchmark. For FLICKR-AES (which provides single scores without label distributions), we replaced the Re-EMD loss with the mean squared error (MSE) loss.

4.2 Training Process

Our entire training process is shown in Fig. 9. Before training begins, we initialize the weights of the MobileNetV2 backbone using ImageNet pretraining as in previous works. The training process consists of three stages. In the first stage, following common practice[7, 14, 31, 54], original images are resized to a fixed resolution of 256×256, randomly cropped to 224×224, and then subjected to random horizontal flipping for data augmentation. This yields an intermediate model. However, considering the possible effects of resizing and cropping on the original images, the model lacks fine-grained details. To address this, we introduce a second stage where we reconstruct the missing information from high-resolution images.

When the training system detects the switching signal in accordance with Eq. (5), the entire training process automatically enters the second stage: mixed-precision training with the Corner Grid data augmentation method that continues until training concludes. Ideally, the model could learn more information from full-resolution images, but our experiment (Fig. 10) and prior work[54] demonstrate that models trained on half-sized input achieve better performance in aesthetic tasks than those trained at full resolution. The image sizes in the AVA dataset vary from 215×160 to 800×800, with an average size of 624×496. Thus, in the second stage, we use half the average size (312×248) as the input size. To maintain the aspect ratio, a constant padding strategy is utilized when the shorter side of an image is less than 312 or 248 pixels.

Upon completion of all second-stage training epochs, the training process advances to the third stage. Considering that the padding regions may confuse the network, we reset the in-

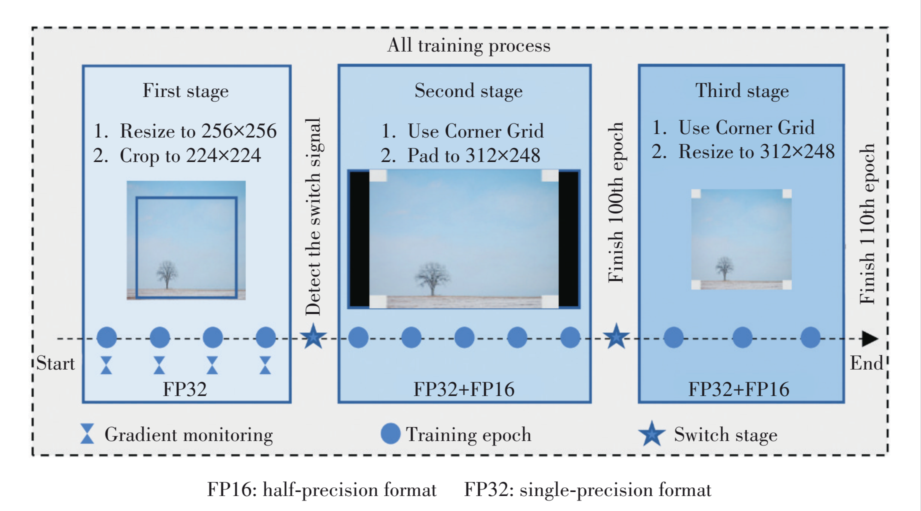
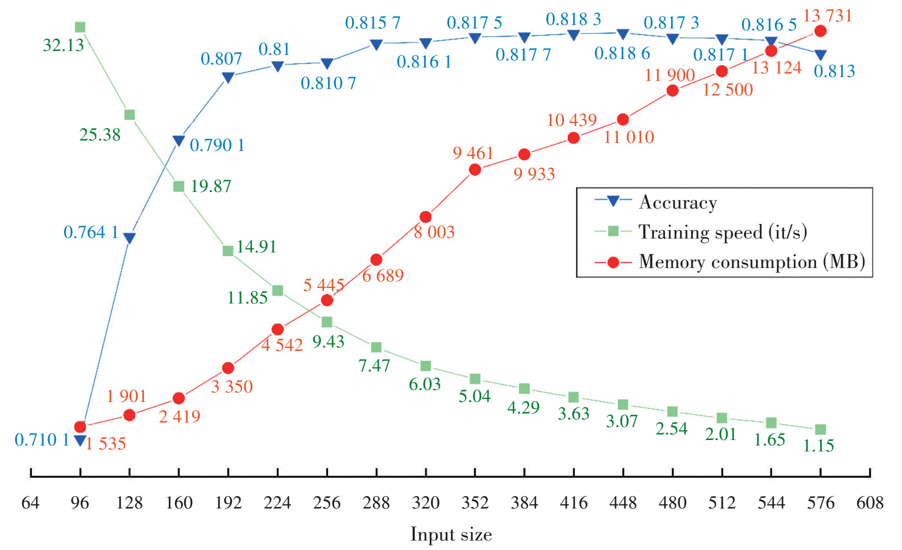


Figure 9. Proposed three-stage training strategy: warm-up (ImageNet→AVA), mixed-precision training (FP32+FP16) with Corner Grid augmentation, and padding refinement



Note: Findings suggest suboptimal performance with both conventional (224×224) and excessively large input sizes for IAA tasks.

Figure 10. Effects of input size variation (AVA dataset, NIMA model) on accuracy, training speed, and memory consumption

put size to 224×224 (without cropping or padding) and conduct rapid mixed-precision retraining for just 10 epochs.

For our Re-EMD loss, we set α=10 and β=0.1, with λ_1=1 and λ_2=0.2 in Eq. (5). The Corner Grid method uses the setting r = 0.1. We use these fixed parameters to make the training more stable compared to learnable alternatives. Our learning rate is fixed at 1e-5 and the Adam optimizer is used, without any decay rate strategy.

4.3 Performance Evaluation

4.3.1 Comparison with SOTA Methods

As seen from Table 1, compared with the 17 SOTA methods on the popular AVA dataset[8], M+MNet achieves the best SRCC (0.770), LCC (0.785), EMD (0.040), and SRCC/Acc ratio (0.934) on AVA with only 4.5 million parameters. This higher ratio indicates that M+MNet better generalizes to the entire range of scores and strikes a good balance between preserving distribution information and increasing discriminability.

We also tested our model on the personalized aesthetic dataset FLICKR-AES. Because our network understands both background and foreground information, it can observably improve the overall performance for personalized aesthetics assessment. As shown in Table 2, our model achieves the best SRCC score of 0.701, surpassing the previous best results by 4.8% SRCC, which means that our model can use a smaller amount of data to learn individual preferences more effectively.

To compare training speeds, we analyze the methods with reported metrics in Table 3. It can be seen that since M+MNet benefits from its lightweight structure as well as its flexible, multi-stage, and mixed-precision training strategy, it is significantly faster than the other comparable methods. This raises a critical question: Can M+MNet enable real-time IAA? Real-time aesthetic guidance for photography/videography is a compelling application. As demonstrated in our released real-time IAA inference video (link), M+MNet achieves 55 fps inference with only 899 MB GPU memory. To the best of our knowledge, this is the first time an IAA model has demonstrated real-time prediction capability, highlighting its potential for mobile deployment to provide real-time interactive guidance. This video also confirms that M+MNet can effectively perceive aesthetics related to the background.

4.3.2 Prediction for Images

Some test images are shown in Fig. 11. Aligning with human cognition, our model assigns higher scores to images that perform better in terms of important aesthetic attributes, such as composition, color, lighting, and depth of field. Because of the incompatible color or unnatural boundary between foreground and background, the corresponding predicted scores are usually lower. Images with low prediction errors (Fig. 11a) usually have both high/low photographic quality and high/low aesthetic quality. However, we also find that the model does not perform well on certain kinds of images (Fig. 11b). These images are generally abstract in their aesthetic expression or gray/black in color, and these kinds of images also appear in fewer numbers in the dataset. In fact, when images do not conform to normal modes of expression in terms of aesthetics and their related attributes, such as color and composition, human evaluators also show inconsistent judgments for the aesthetics of these images, and this is reflected in the lack of uniformity of the opinion labels in the data annotations.

4.4 Ablation Studies

To verify the effectiveness of the various components of the proposed method, we conducted three ablation studies.

4.4.1 Pooling Methods

We conducted experiments with BackNet, ForeNet, and M+MNet using AvgPool, MaxPool, BackPool, and ForePool as the pooling methods. From Table 4, we can observe that the proposed pooling methods consistently improve all metrics. Notably, when applied to NIMA[7], BackPool and ForePool enhance the performance while enabling distinct background/fore-

Table 1. Performance comparison of 18 SOTA IAA models on AVA

Metric		Pub	Code	SRCC ↑	LCC ↑	EMD ↓	Acc ↑	Ratio ↑	Parameter ↓
Code Available 2014 – 2022	RAPID[20]	MM	Lua	0.447*	0.453*	-	0.712	0.628*	**2M**
	AADB[24]	ECCV	Matlab	0.558	0.580*	-	0.773	0.722	8M
	PAM[16]	ICCV	Caffe	0.712*	0.715*	-	0.813*	0.876*	22M
	NIMA[7]	TIP	TF	0.612	0.636	0.050	0.815	0.751	11M
	ALamp[21]	CVPR	Scipy	0.666*	0.671*	-	0.825	0.807*	99M
	MP$_{ada}$[23]	MM	TF	0.727	0.731	-	0.830	0.875	33M
	MLSP[14]	CVPR	TF	0.756	0.757	-	0.817	0.925	24M
	BIAA[34]	TCYB	Torch	0.651*	0.668*	-	0.763*	0.853*	11M
	UIAA[27]	TIP	Matlab	0.719	0.720	0.065	0.808	0.890	23M
	HGCN[28]	CVPR	Jittor	0.665	0.687	0.043	0.846	0.786	44M
Code Not Available 2015 – 2022	DMA[19]	ICCV	N/A	-	-	-	0.754	-	61M
	MNA[51]	CVPR	N/A	-	-	-	0.774	-	138M
	CFAN[52]	IJCAI	N/A	-	-	-	0.810	-	-
	AFDC[29]	CVPR	N/A	0.648	0.671	0.044	0.832	0.779	23M
	PIAA[32]	TIP	N/A	0.677	-	0.047	0.837	0.809	24M
	UGIAA[33]	TMM	N/A	0.692	-	-	**0.851**	0.813	-
	MUSIQ[53]	ICCV	N/A	0.726	0.738	-	-	-	-
Ours		PR	Torch	**0.770**	**0.785**	**0.040**	0.824	**0.934**	4.5M

Note: Models marked with "*" were retrained/re-evaluated using official weights or recommended settings; "-" indicates unavailable metrics (no code/EMD incompatibility).

AADB: Aesthetics and Attributes Database
Acc: accuracy
AFDC: Rating Pictorial Aesthetics Using Deep Learning
ALamp: Adaptive Layout-Aware Multi-Patch Deep Convolutional Neural Network
AVA: Aesthetic Visual Analysis
BIAA: Bilevel Gradient Optimization Image Aesthetics Assessment
CFAN: Cross-domain Feature Aggregation Network
CVPR: Conference on Computer Vision and Pattern Recognition
DMA: Deep Multi-Patch Aggregation
ECCV: European Conference on Computer Vision
EMD: Earth mover's distance
HGCN: Hierarchical Layout-Aware Graph Convolutional Network
IAA: image aesthetics assessment
ICCV: International Conference on Computer Vision
IJCAI: International Joint Conference on Artificial Intelligence
LCC: linear correlation coefficient
MLSP: Multi-Level Spatially Pooled Features
MM: ACM Multimedia
MNA: Multi-Network Aggregation
MPada: Attention-Based Multi-Patch Aggregation
MUSIQ: Multi-Scale Image Quality Transformer
NIMA: Neural Image Assessment
PAM: Personalized Aesthetics Model
PIAA: Personalized Image Aesthetics
PR: Pattern Recognition
RAPID: Rating Pictorial Aesthetics Using Deep Learning
SOTA: state-of-the-art
SRCC: Spearman rank correlation coefficient
TCYB: IEEE Transactions on Cybernetics
TF: TensorFlow
TIP: IEEE Transactions on Image Processing
TMM: IEEE Transactions on Multimedia
UGIAA: Unified Graph-Based Image Aesthetic Assessment
UIAA: Unified Image Aesthetic Assessment

Table 2. Performance comparison of SRCC results of the SOTA models for personalized aesthetics assessment on the FLICKR-AES dataset

Method	10 Images	100 Images
PAM [16]	0.520 ± 0.003	0.553 ± 0.012
PIAA[32]	0.543 ± 0.003	0.639 ± 0.011
UGIAA[33]	0.559 ± 0.002	0.660 ± 0.013
BIAA[34]	0.561 ± 0.005	0.669 ± 0.013
M+MNet	0.585 ± 0.003	0.701 ± 0.009

BIAA: Bilevel Gradient Optimization Image Aesthetics Assessment
M+MNet: Mixed-Precision Multibranch Network
PAM: Personalized Aesthetics Model
PIAA: Personalized Image Aesthetics
SOTA: state-of-the-art
SRCC: Spearman's Rank Correlation Coefficient
UGIAA: Unified Graph-Based Image Aesthetic Assessment

ground feature extraction (Fig. 7), indicating that our proposed methods have better prospects in various IAA models.

4.4.2 Corner Grid

To evaluate the effect of Corner Grid, we selected background-sensitive samples from the AVA dataset, corresponding to 12 000 training images and 3 000 test images. Models lacking robust background perception fail to capture composition guidelines for these images, thus impairing their performance. From Table 5, we can observe that the use of Corner Grid improves the performance (compared with that of M+MNet without Corner Grid) to a certain extent on these background-sensitive samples. To further verify this, we also integrated Corner Grid with NIMA[7], and the results show that Corner Grid also improves the performance of this model, especially its accuracy. Fig. 12 shows that our Corner Grid method can effectively increase the attention area of NIMA. It is worth noting that the proposed pooling methods also improve the prediction performance for background-

Table 3. Comparison of computational costs between M+MNet and reported models (batch size = 16 and input size = 224×224)

Method	Training Speed/(it/s) ↑	Test Speed/(it/s) ↑	GPU Time/ms
NIMA (VGG16)[7]	9.17*	16.09*	85.76
NIMA (Inception)[7]	11.30*	17.64*	39.11
ILGNet[22]	-	-	31.00
NIMA (MobileNet)[7]	15.43*	21.26*	20.23
AFDC (4 patches)[29]	2.08	3.12	-
M+MNet	**34.66**	**90.01**	**13.40**

Note: We used the recommended parameter settings to complete the metrics (*) that are missing in the respective papers; "-" indicates that the metric cannot be obtained.

AFDC: Adaptive Feature Domain Convolution
GPU: Graphics Processing Unit
ILGNet: Integrated Local-Global Network
M+MNet: Mixed-Precision Multibranch Network
NIMA: Neural Image Assessment
VGG16: Visual Geometry Group 16-layer

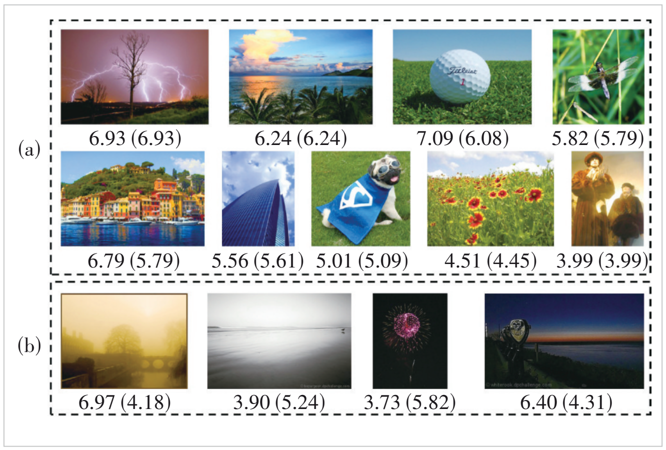


Figure 11. Visualization of images with (a) small and (b) large absolute errors between the ground-truth and predicted (in parentheses) scores

Table 4. Comparison of the proposed and existing pooling methods on the AVA dataset when used in combination with our models and NIMA

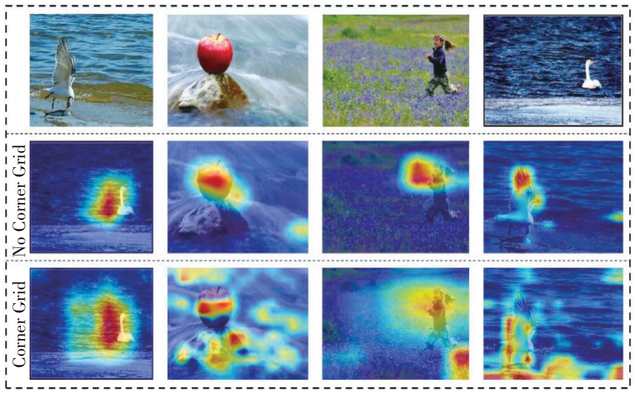
Method	SRCC ↑	LCC ↑	Acc ↑
BackNet (AvgPool)	0.671	0.690	0.789
BackNet (MaxPool)	0.682	0.693	0.786
BackNet (BackPool)	0.723	0.730	0.795
ForeNet (AvgPool)	0.675	0.693	0.786
ForeNet (MaxPool)	0.687	0.699	0.789
ForeNet (ForePool)	0.714	0.732	0.790
M+MNet (AvgPool)	0.716	0.722	0.809
M+MNet (MaxPool)	0.729	0.735	0.810
M+MNet (BackPool)	0.738	0.747	0.813
M+MNet (ForePool)	0.741	0.750	0.819
M+MNet (Fore+BackPool)	**0.770**	**0.785**	**0.824**
NIMA (Original)[7]	0.612	0.636	0.815
NIMA (BackPool)[7]	0.631	0.648	0.820
NIMA (ForePool)[7]	0.635	0.657	0.822

Acc: accuracy
AVA: Aesthetic Visual Analysis
LCC: linear correlation coefficient
M+MNet: Mixed-Precision Multibranch Network
NIMA: Neural Image Assessment
SRCC: Spearman rank correlation coefficient

Table 5. Performance of different architectures on the background-sensitive samples in AVA. We tested our model and NIMA with various pooling methods and Corner Grid

Method	SRCC ↑	LCC ↑	Acc ↑
M+MNet (AvgPool)	0.642	0.649	0.754
M+MNet (MaxPool)	0.634	0.641	0.735
M+MNet (BackPool)	0.670	0.681	0.786
M+MNet (ForePool)	0.665	0.669	0.778
M+MNet (BackPool+ForePool)	0.704	0.716	0.808
M+MNet (BackPool+ForePool+Corner Grid)	**0.739**	**0.744**	**0.814**
NIMA (Original)[7]	0.603	0.620	0.728
NIMA (Corner Grid)[7]	0.611	0.634	0.810

Acc: accuracy
AVA: Aesthetic Visual Analysis
LCC: linear correlation coefficient
M+MNet: Mixed-Precision Multibranch Network
NIMA: Neural Image Assessment
SRCC: Spearman rank correlation coefficient



Note: Activation maps show that our data augmentation method also increases the attention area of this existing model.

Figure 12. Activation maps obtained when using Corner Grid with NIMA

sensitive samples to some extent.

4.4.3 Re-EMD Loss

We used EMD and Re-EMD as the loss functions during training. Table 6 shows that Re-EMD outperforms EMD across all tasks, particularly in ranking metrics (SRCC and LCC). Fig. 13 shows that during the training process, Re-EMD rebalances loss contributions by suppressing noisy samples, enabling the model to focus on more important information. Meanwhile, Re-EMD accelerates convergence of the network relative to EMD. To achieve the best performance shown in Table 6, approximately 200 epochs are needed with the EMD loss, while only 110 epochs are needed with the Re-EMD loss. Furthermore, to verify its generality for various IAA methods, we replaced EMD with Re-EMD in existing works, and the results also show a certain degree of improvement in each metric for these methods.

Table 6. Comparison of the performance achieved by retraining all the IAA models on AVA using the Re-EMD loss in place of the EMD loss

Method	SRCC ↑	LCC ↑	Acc ↑
NIMA (EMD)[7]	0.612	0.636	0.815
NIMA (Re-EMD)[7]	0.633	0.641	0.819
UIAA (EMD)[27]	0.719	0.720	0.808
UIAA (Re-EMD)[27]	0.723	0.731	0.817
HGCN (EMD)[28]	0.665	0.687	**0.846**
HGCN (Re-EMD)[28]	0.689	0.692	0.838
M+MNet (EMD)	0.762	0.766	0.822
M+MNet (Re-EMD)	**0.770**	**0.785**	0.824

Acc: accuracy
EMD: Earth mover's distance
HGCN: Hypergraph Convolutional Network
LCC: linear correlation coefficient
M+MNet: Mixed-Precision Multibranch Network
NIMA: Neural Image Assessment
Re-EMD: rebalanced EMD
SRCC: Spearman rank correlation coefficient
UIAA: Unified Image Aesthetic Assessment

5 Discussion

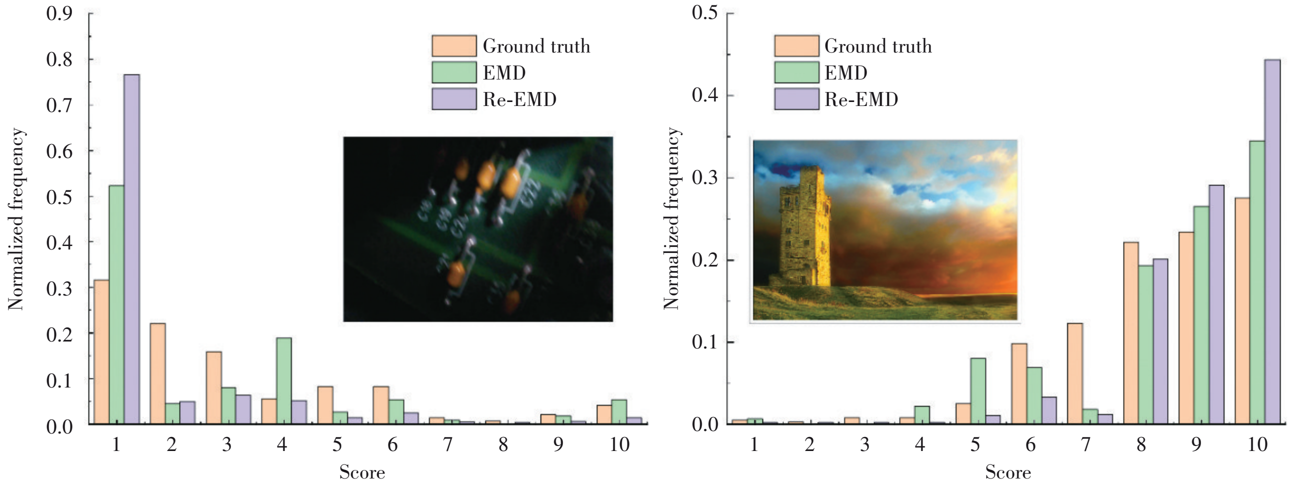
In the field of IAA, it is common for IAA models to experience reduced accuracy when evaluating dark scenes. This issue can be understood and improved from several perspectives as follows. 1) lighting conditions and contrast: Dark scenes often suffer from insufficient lighting, leading to low image contrast and loss of detail. Under low-light conditions, the increase in image noise can also impact the accuracy of aesthetic assessments. 2) Bias in training dataset: The existing IAA datasets used for training have a limited number of dark scene samples, limiting the model's ability to understand and evaluate these types of scenes. The model ' s performance largely depends on the diversity and quality of its training data. 3) Feature extraction capability: The details and texture features in dark scenes might not be as rich as in brighter scenes, making it difficult for the model to extract and utilize these features accurately for evaluation.

To improve the model's evaluation accuracy in dark scenes, we consider the following changes in future work:

1) Enhancing the training dataset: We can add more high-quality dark scene images to the training dataset to improve the model's performance in processing these images.

2) Adopting specialized network architectures: We will develop neural network structures optimized for low-light conditions, such as convolutional networks with enhanced light perception capabilities.

3) Conducting multimodal learning: We will combine other information about the image, such as metadata and contextual scene information, to assist in the aesthetic assessment of dark scenes.



EMD: Earth mover's distance Re-EMD: rebalanced EMD

Figure 13. Results of normalizing (0 – 1) distributions of the ground truth and losses contributed by each score based on EMD and Re-EMD losses during training

6 Conclusions and Future Work

In this paper, we show that enhancing the attention to background information in CNN-based models can effectively improve performance on IAA tasks. We introduce the M+MNet model and use a mixed-precision approach in our multi-stage training strategy while proposing a novel Re-EMD loss function to boost performance. The results suggest that our method not only achieves SOTA performance on all IAA tasks but also enables much faster training with reduced training costs. The proposed data augmentation method, Corner Grid, successfully directs more model attention to background areas, though its full performance potential remains to be explored. Our proposals can be independently implemented in combination with existing methods to overcome the main stumbling blocks for IAA tasks. The commercial application of IAA models faces several technical challenges, particularly from the perspective of their "black box" nature, which refers to the difficulty in understanding and interpreting how these models make decisions. As part of future work, we will further explore methods that can help models understand aesthetics while designing explainable IAA models.

References

[1] ITTI L, KOCH C. Computational modelling of visual attention [J]. Nature reviews neuroscience, 2001, 2(3): 194 – 203. DOI: 10.1038/35058500

[2] SIAGIAN C, ITTI L. Rapid biologically-inspired scene classification using features shared with visual attention [J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 29(2): 300 – 312. DOI: 10.1109/TPAMI.2007.40

[3] BIEDERMAN I. Do background depth gradients facilitate object identification? [J]. Perception, 1981, 10(5): 573 – 578. DOI: 10.1068/p100573

[4] POTTER M C. Meaning in visual search [J]. Science, 1975, 187(4180): 965 – 966. DOI: 10.1126/science.1145183

[5] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization [C]//Proc. IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 618 – 626. DOI: 10.1109/ICCV.2017.74

[6] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database [C]//Proc. IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 248 – 255. DOI: 10.1109/CVPR.2009.5206848

[7] TALEBI H, MILANFAR P. NIMA: neural image assessment [J]. IEEE transactions on image processing, 2018, 27(8): 3998 – 4011. DOI: 10.1109/TIP.2018.2831899

[8] MURRAY N, MARCHESOTTI L, PERRONNIN F. AVA: a large-scale database for aesthetic visual analysis [C]//Proc. IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 2408 – 2415. DOI: 10.1109/CVPR.2012.6247954

[9] GAO Z T, WANG L M, WU G S. LIP: local importance-based pooling [C]//Proc. IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 3355 – 3364. DOI: 10.1109/iccv.2019.00345

[10] ZHONG Z, ZHENG L, KANG G L, et al. Random erasing data augmentation [C]//Proc. AAAI Conference on Artificial Intelligence. AAAI, 2020: 13001 – 13008. DOI: 10.1609/aaai.v34i07.7000

[11] DEVRIES T, TAYLOR G W. Improved regularization of convolutional neural networks with cutout [EB/OL]. (2017-08-15)[2024-01-23]. https://arxiv.org/abs/1708.04552

[12] SINGH K K, LEE Y J. Hide-and-seek: forcing a network to be meticulous for weakly-supervised object and action localization [C]//Proc. IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 3544 – 3553. DOI: 10.1109/ICCV.2017.381

[13] HE K M, CHEN X L, XIE S N, et al. Masked autoencoders are scalable vision learners [EB/OL]. (2021-11-11)[2024-03-20]. https://arxiv.org/abs/2111.06377

[14] HOSU V, GOLDLUCKE B, SAUPE D. Effective aesthetics prediction with multi-level spatially pooled features [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 9375 – 9383. DOI: 10.1109/cvpr.2019.00960

[15] MICIKEVICIUS P, NARANG S, ALBEN J, et al. Mixed precision training [C]//Proc. International Conference on Learning Representations (ICLR). OpenReview, 2018. DOI: 10.48550/arXiv.1710.03740

[16] REN J, SHEN X H, LIN Z, et al. Personalized image aesthetics [C]//Proc. IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 638 – 647. DOI: 10.1109/ICCV.2017.76

[17] LUO Y W, TANG X O. Photo and video quality evaluation: focusing on the subject [C]//European Conference on Computer Vision. ECCV, 2008: 386 – 399. DOI: 10.1007/978-3-540-88690-7_29

[18] DATTA R, JOSHI D, LI J, et al. Studying aesthetics in photographic images using a computational approach [C]//European Conference on Computer Vision. ECCV, 2006: 288 – 301. DOI: 10.1007/11744078_23

[19] LU X, LIN Z, SHEN X H, et al. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation [C]//Proc. IEEE International Conference on Computer Vision (ICCV). IEEE, 2015: 990 – 998. DOI: 10.1109/ICCV.2015.119

[20] LU X, LIN Z L, JIN H L, et al. RAPID: rating pictorial aesthetics using deep learning [C]//Proc. ACM International Conference on Multimedia. ACM, 2014: 457 – 466. DOI: 10.1145/2647868.2654926

[21] MA S, LIU J, CHEN C W. A-lamp: adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment [C]//Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 722 – 731. DOI: 10.1109/CVPR.2017.84

[22] JIN X, WU L, LI X D, et al. ILGNet: inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation [J]. IET computer vision, 2019, 13(2): 206 – 212. DOI: 10.1049/iet-cvi.2018.5249

[23] SHENG K K, DONG W M, MA C Y, et al. Attention-based multi-patch aggregation for image aesthetic assessment [C]//Proc. 26th ACM International Conference on Multimedia. ACM, 2018: 879 – 886. DOI: 10.1145/3240508.3240554

[24] KONG S, SHEN X H, LIN Z, et al. Photo aesthetics ranking network with attributes and content adaptation [C]// European Conference on Computer Vision. ECCV, 2016: 662 – 679. DOI: 10.1007/978-3-319-46448-0_40

[25] ZHANG X D, GAO X B, LU W, et al. Beyond vision: a multimodal recurrent attention convolutional neural network for unified image aesthetic prediction tasks [J]. IEEE transactions on multimedia, 2020, 23: 611 – 623. DOI: 10.1109/TMM.2020.2985526

[26] HOU L, YU C P, SAMARAS D. Squared earth mover's distance-based loss for training deep neural networks [EB/OL]. (2016-11-18)[2024-03-19]. https://arxiv.org/abs/1611.05916

[27] ZENG H, CAO Z S, ZHANG L, et al. A unified probabilistic formulation of image aesthetic assessment [J]. IEEE transactions on image processing, 2019, 29: 1548 – 1561. DOI: 10.1109/TIP.2019.2941778

[28] SHE D Y, LAI Y K, YI G X, et al. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 8475 – 8484. DOI: 10.1109/cvpr46437.2021.00837

[29] CHEN Q Y, ZHANG W, ZHOU N, et al. Adaptive fractional dilated convolution network for image aesthetics assessment [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE,

2020: 14114 – 14123. DOI: 10.1109/cvpr42600.2020.01412

[30] MURRAY N, GORDO A. A deep architecture for unified aesthetic prediction [EB/OL]. (2017-08-16)[2024-03-19]. https://arxiv.org/abs/1708.04890

[31] ZHAO L, SHANG M M, GAO F, et al. Representation learning of image composition for aesthetic prediction [J]. Computer vision and image understanding, 2020, 199: 103024. DOI: 10.1016/j.cviu.2020.103024

[32] LI L D, ZHU H C, ZHAO S C, et al. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment [J]. IEEE transactions on image processing, 2020, 29: 3898 – 3910. DOI: 10.1109/TIP.2020.2968285

[33] LYU P, FAN J, NIE X, et al. User-guided personalized image aesthetic assessment based on deep reinforcement learning [EB/OL]. (2021-06-14)[2024-03-19]. https://arxiv.org/abs/2106.07488

[34] ZHU H C, LI L D, WU J J, et al. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization [J]. IEEE transactions on cybernetics, 2022, 52(3): 1798 – 1811. DOI: 10.1109/TCYB.2020.2984670

[35] ZHANG X D, GAO X B, LU W, et al. A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction [J]. IEEE transactions on multimedia, 2019, 21(11): 2815 – 2826. DOI: 10.1109/TMM.2019.2911428

[36] WANG W S, YANG S, ZHANG W S, et al. Neural aesthetic image reviewer [J]. IET computer vision, 2019, 13(8): 749 – 758. DOI: 10.1049/iet-cvi.2019.0361

[37] SAKURIKAR P, MEHTA I, BALASUBRAMANIAN V N, et al. RefocusGAN: scene refocusing using a single image [C]//European Conference on Computer Vision. ECCV, 2018: 519 – 535. DOI: 10.1007/978-3-030-01225-0_31

[38] SITZMANN V, DIAMOND S, PENG Y F, et al. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging [J]. ACM transactions on graphics, 2018, 37(4): 1 – 13. DOI: 10.1145/3197517.3201333

[39] PURI R, KIRBY R, YAKOVENKO N, et al. Large scale language modeling: converging on 40GB of text in four hours [C]//Proc. IEEE International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD). IEEE, 2018: 290 – 297. DOI: 10.1109/SBAC-PAD.2018.00043

[40] COURBARIAUX M, BENGIO Y, DAVID J P. BinaryConnect: training deep neural networks with binary weights during propagations [C]//Proc. Advances in Neural Information Processing Systems (NeurIPS). Curran Associates, 2015: 3123 – 3131. DOI: 10.48550/arXiv.1511.00363

[41] HUBARA I, COURBARIAUX M, SOUDY D, et al. Quantized neural networks: training neural networks with low precision weights and activations [J]. Journal of machine learning research, 2017, 18(1): 6869 – 6898. DOI: 10.5555/3122009.3242014

[42] HE Q, WEN H, ZHOU S, et al. Effective quantization methods for recurrent neural networks [EB/OL]. (2016-11-30)[2024-03-19]. https://arxiv.org/abs/1611.10176

[43] ZHOU S, WU Y, NI Z, et al. DoReFa-Net: training low bitwidth convolutional neural networks with low bitwidth gradients [EB/OL]. (2016-06-20)[2024-03-19]. https://arxiv.org/abs/1606.06160

[44] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 4510 – 4520. DOI: 10.1109/CVPR.2018.00474

[45] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proc. Advances in Neural Information Processing Systems (NeurIPS). Curran Associates, 2017: 5998 – 6008. DOI: 10.5555/3295222.3295349

[46] ZHANG H, GOODFELLOW I J, METAXAS D N, et al. Self-attention generative adversarial networks [C]//Proc. International Conference on Machine Learning (ICML). PMLR, 2018: 7354 – 7363. DOI: 10.5555/3327757.3360384

[47] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [C]//European Conference on Computer Vision. ECCV, 2018: 833 – 851. DOI: 10.1007/978-3-030-01234-2_49

[48] LIN M, CHEN Q, YAN S. Network in network [C]//Proc. International Conference on Learning Representations (ICLR). ICLR, 2014: 1 – 10. DOI: 10.48550/arXiv.1312.4400.

[49] LIU D, PURI R, KAMATH N, et al. Composition-aware image aesthetics assessment [C]//Proc. Winter Conference on Applications of Computer Vision (WACV). IEEE, 2020: 3569 – 3578. DOI: 10.1109/WACV45572.2020.9093626

[50] BUCHSBAUM G. A spatial processor model for object colour perception [J]. Journal of the franklin institute, 1980, 310(1): 1 – 26. DOI: 10.1016/0016-0032(80)90058-7

[51] MAI L, JIN H L, LIU F. Composition-preserving deep photo aesthetics assessment [C]//Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 497 – 506. DOI: 10.1109/CVPR.2016.60

[52] WANG G L, YAN J C, QIN Z. Collaborative and attentive learning for personalized image aesthetic assessment [C]//Proc. Twenty-Seventh International Joint Conference on Artificial Intelligence. IJCAI, 2018: 957 – 963. DOI: 10.24963/ijcai.2018/133

[53] KE J J, WANG Q F, WANG Y L, et al. MUSIQ: multi-scale image quality transformer [C]//Proc. IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 5128 – 5137. DOI: 10.1109/ICCV48922.2021.00510

[54] HOSU V, LIN H H, SZIRANYI T, et al. KonIQ-10k: an ecologically valid database for deep learning of blind image quality assessment [J]. IEEE transactions on image processing, 2020, 29: 4041 – 4056. DOI: 10.1109/TIP.2020.2967829

Biographies

HE Shuai is a postdoctoral researcher in computer science at Beijing University of Posts and Telecommunications (BUPT), China. His research interests include image processing and image aesthetics assessment.

LIU Limin is pursuing a master's degree in computer science at Beijing University of Posts and Telecommunications (BUPT), China, with a research focus on image processing and image aesthetics assessment.

WANG Zhanli is an engineer at ZTE Corporation, with a research focus on image processing.

LI Jinliang is an engineer at ZTE Corporation, with a research focus on image processing.

MAO Xiaojun is an engineer at ZTE Corporation, with a research focus on image processing and image quality assessment.

MING Anlong (mal@bupt.edu.cn) received his PhD from Beijing University of Posts and Telecommunications (BUPT), China in 2008. He is currently a professor with the School of Computer Science, BUPT. His research interests include computer vision and robot vision.