Pseudo-Labeling and Meta Reweighting Learning for Image Aesthetic Quality Assessment

Xin Jin, Hao Lou, Heng Huang, Xinning Li, Xiaodong Li, Shuai Cui, Xiaokun Zhang, and Xiqiao Li

Abstract—In the tasks of image aesthetic quality assessment, it is difficult to reach both the high score area and low score area due to the normal distribution of aesthetic datasets. To reduce the error in labelling and solve the problem of normal data distribution, we propose a new aesthetic mixed dataset with classification and regression called AMD-CR, and we train a meta reweighting network to reweight the loss of training data differently. In addition, we provide a training strategy according to different stages based on pseudo labels, and then we use it for aesthetic training according to different stages in classification and regression tasks. In the construction of the network structure, we construct an aesthetic adaptive block (AAB) structure that can adapt to any size of the input images. Besides, we also use the efficient channel attention (ECA) to strengthen the feature extracting ability of each task. The experimental result shows that our method improves 0.1112 compared with the conventional method in SROCC. The method can also help to find best aesthetic path planning for unmanned aerial vehicles (UAV) and vehicles.

Index Terms—Image aesthetic quality assessment, meta reweighting network, pseudo labels, aesthetic adaptive block, path planning.

I. INTRODUCTION

AESTHETICS was raised by the German philosopher Baumgarten in 1750. It focuses on the aesthetic relationship between the public and the world. The aesthetic research subjects are the aesthetic activities with strong subjectivity. Only the experts can evaluate the abstractly aesthetic consensus in real life. However, the public can not give the accurate aesthetic evaluation without professional training, which leads to computational aesthetics [1].

The goal of computational aesthetics is aesthetic intelligence that computers and robots are able to recognize, produce, and create beauty. That can make computers and robots focus on human cognition and the inner mechanism of recognizing beauty. Computational visual aesthetics [2] is an intelligent comprehension about visual information by visual computation techniques. Nowadays, visual computation aesthetics depends on deep learning. In the related researches, the visual computation aesthetics can get neural network modal mainly by training massive data. So, the model can give the evaluation of aesthetic quality. Although the generally existing aesthetic benchmark dataset like AVA [3], AADB [4], PCCD [5] and others have numeric score labels, most of them are not over 20,000 images. AVA is the largest dataset with 25,553 images, but the overall scores of over 7 and under 3 only take 4.4%. In the training process, that can trigger seriously underfitting problems. AADB includes 9,958 images, and each image has 11 aesthetic attribute scores. However, the image amount is too few and the aesthetic labels don’t have good labelling quality. CHUKPQ [6] only has the simple two-value labels instead of the numeric labels. The above datasets have obvious disadvantages. Moreover, comparing with target detection, semantic segmentation, semantic classification, and other tasks, it is more abstract to extract aesthetic features in detail. What’s harder is to extract lighting, color, composition, and other attribute feature of aesthetics. These bring a lot of challenges to evaluation of aesthetic quality.

In order to solve those problems, we propose a method of aesthetic quality evaluation based on pseudo-labelling and meta-learning. Specifically, we make the following four contributions:

1. We filter and construct a dataset called AMD-CR with an appropriate scale. AMD-CR has both more reasonable data distribution and better aesthetic quality labels than the classic aesthetic datasets.
2. We introduce meta-learning in aesthetic sample re-weight. We use meta reweighting network to provide appropriate weights of the training samples.
3. We propose the aesthetic adaptive block (AAB) to compose aesthetic network model and use it to adapt to images with any ratio between length and width.
4. To the best of our knowledge, we are the first to use pseudo-labelling in the aesthetic tasks. Our method shows...
good results in image aesthetic quality assessment and multi-task network [20]. These methods quantize aesthetic evaluation by contrasting composition, light, color, matching, exposure, depth of field and camera use. However, all researches are limited by the amount of data and data distribution. Our research [9] has great improvement in binary classification of aesthetic evaluation and numeric evaluation of regression tasks. Beside the application of basic aesthetic evaluation, we propose a multistream network, composed by spatial, motion, and structural streams, to gain the multimodal features of path planning. That inspires us to explore path planning based on image aesthetic quality assessment.

C. Meta Learning

In deep learning, the model has to be retrained when scenes change. Meta learning, learn to learn, can learn the ability of human learning by limit data [21]. Meta learning mainly includes memory [22], prediction of gradient [23], attention mechanisms [24], LSTM [25], reinforcement learning [26] and others. By the inspiration of meta learning [21], [25], the recent researches study reweight cases adaptively from data samples. That shows learning is more automatic and dependent. This idea can effectively solve the long-tail problem caused by few data tags.

III. RESEARCH METHODOLOGY

In this section, we mainly introduce the structure and the algorithm of the aesthetic assessment model in detail. The network structure is constructed in Fig.1.

A. Meta Reweighting Network

The loss functions of conventional deep learning are generally cross entropy or mean square error loss function, which calculate the loss of average on batch size samples. These methods are easily influenced by the sample distribution. In the classification task, if the proportion of the positive samples is high, then the final transmission of the loss value must be influenced by most of the positive sample loss. Although a few negative samples have higher loss, the network will be less influenced by these negative samples. The idea of the meta reweighting in this paper is to learn the loss weight of the data automatically, optimize the normal data distribution, and enhance the model robustness.

We apply the method of meta reweighting mentioned in [27] to aesthetic quality assessment tasks for the first time. The idea of meta reweighting is to minimize the loss

\[
\frac{1}{N} \sum_{i=1}^{N} \ell (y_i, f(x_i))
\]

in the training set and extract the best classifier parameters \( w^* \). In the biased training set, each sample weight is expressed as

\[
\mathcal{V} \left( L_i^{\text{train}} (w); \Theta \right)
\]

where \( L_i^{\text{train}} (w) = \ell (y_i, f(x_i)) \) means the loss of \( i \)th training sample; \( \Theta \) represents the parameter of the meta reweighting network(MRN). The best parameter \( w^* \) is calculated as follows:

\[
w^*(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{V} \left( L_i^{\text{train}} (w); \Theta \right) L_i^{\text{train}} (w)
\]

In order to learn hyperparametric \( \Theta \) automatically, we generate a multilayer perceptron based on the hidden layers in
the method of meta learning, shown as the yellow network in Fig. 1. The input of the meta reweighting network is the loss of samples, and the output is the weight of the loss of samples. We extract a few high quality images from the training set $S_{\text{train}}$ to construct the meta set $S_{\text{meta}}$. $N$ is the amount of $S_{\text{train}}$, and $M$ is the amount of $S_{\text{meta}}$, $N \gg M$. The advantages of $S_{\text{meta}}$ are that the data is correctly labeled and equal in distribution. According to the score labels of aesthetic quality assessment datasets, we divide the data into 10 segments equally and extract 200 images from each segment artificially. If the amount of any segment is less than 200, we will choose appropriate amount of images from the adjacent segments to replenish the lack images. Finally, we filter 2000 images to construct $S_{\text{meta}}$. We update the $\Theta$ and $w$ through a single optimization cycle. There are two stages in training process. We firstly update the parameter $\Theta$ of the MRN, and then update the parameter $w$ of the main network in each iteration optimization. The update equation for the first stage is defined as follows:

$$
\Theta^{(t+1)} = \Theta^{(t)} - \beta \frac{1}{m} \sum_{i=1}^{m} \nabla_{\Theta} L_{i}^{\text{meta}} \left( \hat{w}^{(t)}(\Theta) \right)_{w^{(t)}}
$$

(2)

$\beta$ is the step size of the MRN. $\nabla_{\Theta} L_{i}^{\text{meta}} \left( \hat{w}^{(t)}(\Theta) \right)_{w^{(t)}}$ represents the gradient of the MRN on meta set, $\hat{w}^{(t)}(\Theta)$ is calculated as follows:

$$
\hat{w}^{(t)}(\theta) = w^{(t)} - \alpha \sum_{i=1}^{n} v \left( L_{i}^{\text{train}}(w^{(t)}); \Theta \right) \nabla_{w} L_{i}^{\text{train}}(w)_{w^{(t)}}
$$

(3)

$\alpha$ is the step size of the main network. In the first stage, we need to calculate the weighted loss of $S_{\text{train}}$ on the main network. In order to distinguish real training from meta-learning training, we copy the parameters of the main network at each single cycle training. The parameters of MRN are optimized by Equation (2).

In the second stage, we optimize and update the main network according to the loss of $S_{\text{train}}$ and the weight from MRN on the meta network as follows:

$$
\hat{w}^{(t+1)} = w^{(t)} - \alpha \sum_{i=1}^{n} v \left( L_{i}^{\text{train}}(w^{(t)}); \Theta^{(t+1)} \right) \nabla_{w} L_{i}^{\text{train}}(w)_{w^{(t)}}
$$

(4)

$\beta$ is the step size of the MRN. $\nabla_{\Theta} L_{i}^{\text{meta}} \left( \hat{w}^{(t)}(\Theta) \right)_{w^{(t)}}$ represents the gradient of the MRN on meta set, $\hat{w}^{(t)}(\Theta)$ is calculated as follows:

$$
\hat{w}^{(t)}(\theta) = w^{(t)} - \alpha \sum_{i=1}^{n} v \left( L_{i}^{\text{train}}(w^{(t)}); \Theta \right) \nabla_{w} L_{i}^{\text{train}}(w)_{w^{(t)}}
$$

(3)

$\alpha$ is the step size of the main network. In the first stage, we need to calculate the weighted loss of $S_{\text{train}}$ on the main network. In order to distinguish real training from meta-learning training, we copy the parameters of the main network at each single cycle training. The parameters of MRN are optimized by Equation (2).

In the second stage, we optimize and update the main network according to the loss of $S_{\text{train}}$ and the weight from MRN on the meta network as follows:

$$
\hat{w}^{(t+1)} = w^{(t)} - \alpha \sum_{i=1}^{n} v \left( L_{i}^{\text{train}}(w^{(t)}); \Theta^{(t+1)} \right) \nabla_{w} L_{i}^{\text{train}}(w)_{w^{(t)}}
$$

(4)

$\beta$ is the step size of the MRN. $\nabla_{\Theta} L_{i}^{\text{meta}} \left( \hat{w}^{(t)}(\Theta) \right)_{w^{(t)}}$ represents the gradient of the MRN on meta set, $\hat{w}^{(t)}(\Theta)$ is calculated as follows:

$$
\hat{w}^{(t)}(\theta) = w^{(t)} - \alpha \sum_{i=1}^{n} v \left( L_{i}^{\text{train}}(w^{(t)}); \Theta \right) \nabla_{w} L_{i}^{\text{train}}(w)_{w^{(t)}}
$$

(3)

$\alpha$ is the step size of the main network. In the first stage, we need to calculate the weighted loss of $S_{\text{train}}$ on the main network. In order to distinguish real training from meta-learning training, we copy the parameters of the main network at each single cycle training. The parameters of MRN are optimized by Equation (2).

In the second stage, we optimize and update the main network according to the loss of $S_{\text{train}}$ and the weight from MRN on the meta network as follows:

$$
\hat{w}^{(t+1)} = w^{(t)} - \alpha \sum_{i=1}^{n} v \left( L_{i}^{\text{train}}(w^{(t)}); \Theta^{(t+1)} \right) \nabla_{w} L_{i}^{\text{train}}(w)_{w^{(t)}}
$$

(4)
resizing. In AAB, we adjust the long edge of the image to 800 proportionally, and then we fill the short edge with 0 to make it reach 800. So, we construct an $800 \times 800$ input image. Then, for the first convolution operation of the EfficientNet-B4, we use an adaptive average pooling layer to reshape the feature map of $48 \times 190 \times 190$; 48 is the amount of the feature channel. Finally, the feature map enters the remaining network layers in EfficientNet-B4. Three methods are shown as Fig.2 and Fig.3.

C. Efficient Channel Attention

Recently, the channel attention mechanism [28] shows a great potential in improving the performance of deep learning network models. However, most of the previous focused on building complex attention modules to achieve good performance. In order to reduce the complexity of models, this paper cites the efficient channel attention (ECA) module [7]. It adopts a local cross-channel attention mechanism and the adaptive one-dimensional convolution kernel, which can improve performance obviously.

The ECA captures local cross-channel attention information by interacting with each channel and its $k$ neighbors. This can be effectively implemented by a 1D convolution with the kernel size of $k$, where $k$ represents the local cross-channel attention and means the number of neighbor nodes involved in the attention prediction of this channel. In order to avoid adjusting $k$ in the validation phase, the ECA proposes a method that $k$ can be adaptively proportional to the channel dimensions, $k$ is expressed as follows:

$$k = \gamma(C) = \left\lfloor \frac{\log_2(C) + b}{\gamma} \right\rfloor \text{odd}$$

$|r|_{\text{odd}}$ means the nearest odd number of $r$. The value of $\gamma$ is 2, and the value of $b$ is 1. We can get $k = 7$ if $C = 1792$.

D. Multitask Network

Both the aesthetic multi-classification and regression tasks share the same main network parameters to extract features. The multi-classification network and the regression network structure are similar as shown in Fig.1. The main network outputs the feature map as $N \times 1792 \times 11 \times 11$. $N$ is the batch size. The feature map will be transformed as $N \times 1792$ by a global average pooling layer, and then it will be transformed into two fully connected layers. After the second fully connected layer, the output will be 10 nodes in the multi-classification network and 1 node in the regression network. Loss function formula of multi-classification network is expressed as follows:

$$Loss_{\text{class}} = \sum_x p(x) \log q(x)$$

In this equation, $p(x)$ represents the predicted output probability, and $q(x)$ represents the ground-truth probability. The probability is obtained as follows:

$$y_i = \frac{e^{x_i}}{\sum_{j \in 1} e^{x_j}}$$

F. Datasets

Both the aesthetic multi-classification and regression tasks share the same main network parameters to extract features. The multi-classification network and the regression network structure are similar as shown in Fig.1. The main network outputs the feature map as $N \times 1792 \times 11 \times 11$. $N$ is the batch size. The feature map will be transformed as $N \times 1792$ by a global average pooling layer, and then it will be transformed into two fully connected layers. After the second fully connected layer, the output will be 10 nodes in the multi-classification network and 1 node in the regression network. Loss function formula of multi-classification network is expressed as follows:

$$Loss_{\text{regre}} = \frac{1}{N} \sum_i (s_i - \hat{s}_i)^2$$

$N$ is the number of samples. $s_i$ is the output score, and $\hat{s}_i$ is the ground-truth score.

E. Piecewise Strategy Based on Pseudo-Labeling of the Binary Classification Task

The pseudo-labelling was firstly proposed in the [29], which is not completely consistent with the ground-truth labels; and the ground-truth labels are often used as the hard labels. Pseudo labels can be considered as soft labels after training. We propose a piecewise training method based on pseudo labels of the binary classification task. For first stage, we train a binary classifier, and the images are divided into 0 and 1 based on the boundary of 5 points, while 0 represents low aesthetic quality and 1 represents high aesthetic quality. TABLE I shows the accuracy and error rates of each segment in the testing set.

Although there are still a few false classification samples in 2.0-3.0 and 8.0-9.0 intervals, the error rate in the middle interval 4.0-6.0 is higher; it is because, in the middle interval, there are no huge difference in the scores of images in all perspectives of aesthetic quality assessment. However, this does not influence the final assessment of image aesthetics by the network model. Through continuous score regression, the assessment measures of the model can be measured by quantitative errors. The results of the classification task by the binary classifier can be regarded as pseudo labels of the model. In the second stage, the pseudo labels redivide the datasets into two datasets. We performed multi-classification training and fine-grained regression training separately to extract better image aesthetic features. The experimental results in Section IV show that the piecewise strategy helps to improve the accuracy of the regression.

F. Data

In order to construct an aesthetic dataset with reasonable distribution and high quality of aesthetic annotation, we filter and reconstruct a dataset called AMD-CR. There are two
datasets in AMD-CR: aesthetic mixed dataset with classification (AMD-C) applied to the binary classification training task, and aesthetic mixed dataset with regression (AMD-R) applied to the regression task.

1) Construction of AMD-C: The part of AMD-C comes from image aesthetic benchmark datasets: DPChallenge, Photo.net, AVA [3], CUHKPQ [6] and SPAQ [30], AVA and DPChallenge come from www.dpchallenge.com. Photo.net is from www.photo.net; another part comes from self-built datasets: GLAMOUR, OUTDOOR, NG and PSA. GLAMOUR is from www.glamour-photos.org, OUTDOOR is from www.outdoor-photos.com, NG is from www.dili360.com, PSA is from www.paschini.org. We collected the high-quality data by downloading high-quality on the professional photography websites. We obtain a portion of the public images from these websites and integrate them into our aesthetic dataset. We distribute continuous labels to [0, 10.0] through standardized fractional labelling and map binary labels to 0 and 1. 0 represents the images with low aesthetic quality and 1 represents the images with high aesthetic quality. Discrete datasets only have binary labels, while continuous datasets both have binary labels and continuous labels. We focused on selecting images of high and low quality, and we performed a preliminary selection of the images with general aesthetic quality to balance the data distribution. TABLE II shows the constitution of AMD-C.

2) Construction of AMD-R: We filter the data with the continuous labels in AMD-CR. The images of low and high segments with the scores less than 4.0 points or more than 6.0 points are reserved, and the middle segment images (4.0-6.0) are randomly sampled. Then we get 59,371 images from AMD-R. The segment distribution of AMD-R is shown in Fig. 4.

3) Construction of AMD-C: We removed the images ranging between 4.0 points and 6.0 points in the AMD-CR to ensure that the ratio of positive and negative samples is 1:1. Finally, we obtain AMD-C containing 61,660 images.

### IV. EXPERIMENT

#### A. Training Details

We set the classification batch size to be 32, the regression batch size to be 64 and the learning rate to be 0.0001. We use Adam as the optimizer; betas are set as (0.98, 0.999); weight decay is set as 0.0001. If the accuracy rate of classification is not improved in two consecutive rounds, or the regression loss is not decreased in two consecutive rounds, the learning rate will multiply by 0.5. Our running environment is in Pytorch 1.5.0 and Nvidia TITAN XPs.

The datasets used in this paper are AMD-C and AMD-R. We divide the dataset into three sets; the ratio of training set and validation set and testing set is 8:1:1. EfficientNet-B4 is the main network in the model. After we firstly trained the binary classifier, we will get the training model as $C_2$. Then the training set and validation set are divided into $S_{train0}$, $S_{train1}$, $S_{valid0}$, and $S_{valid1}$ according to the classifier, 0 represents the low quality and 1 represents the high quality in aesthetics. Then we use $S_{train0}$, $S_{valid0}$ and $S_{train1}$, $S_{valid1}$ to train classification and regression models for the next stage.

In the aesthetic assessment training, we start with a 10-class classification training, with [0,10.0] divided into ten average segments, corresponding to the category A when the scoring label is in the $(A, A+1.0]$ interval, $A \in \{0,1,2,3,4,5,6,7,8,9\}$. If there is an image with 0 score, it is classified as the 0 category. When we perform the ten-class training task with the images of the same pseudo labels obtained by the binary classifier, the parameters of the regression network should not be returned. We use cross-entropy function to get classification loss. Then, we release the regression network and freeze the main network, classification network, and MRN to train the regression model, we can get models $R_0$ and $R_1$ corresponding to the $S_{train0}$ and $S_{train1}$ respectively. In the same method, we can get regression model $R_{valid}$ training on all the AMD-R. Finally we can get the result score like this: if $C_2(i) = 0$, $score = 0.5 * \frac{R_0(i) + R_{valid}(i)}{2}$; else $score = 0.5 * \frac{R_1(i) + R_{valid}(i)}{2}$, $i \in S_{test}$.

#### B. Analysis of Experimental Results

We use these indicators for assessment: first, we calculate mean square error (MSE): calculating the loss between the output score and the ground-truth score. It is as follows:

$$MSE = \frac{1}{N} \sum_{i} (r_i - \hat{r_i})^2$$  \hspace{1cm} (9)

$N$ is the number of samples. $r_i$ is the output score, and $\hat{r_i}$ is the ground-truth score.

Mean absolute error (MAE) measures the differences between ground-truth and regressed scores directly, and then
Fig. 5. Test samples. G for the ground-truth score, M for the score of MRN+AAB+ECA(PCR), A for the score of AAB+ECA(PCR), R for the score of Resizing+ECA(CR) and C for the score of Cropping+ECA(CR).

find the mean of differences. It is as follows:

\[ MAE = \frac{1}{N} \sum_i |r_i - \hat{r}_i| \quad (10) \]

\( N \) is the number of samples, \( r_i \) is the output score, and \( \hat{r}_i \) is the ground-truth score.

We calculate Spearman’s rank order correlation coefficient (SROCC): the correlation between the output score and the ground-truth score. It is as follows:

\[ SROCC = 1 - \frac{6 \sum_i (r_i - \hat{r}_i)^2}{N^3 - N} \quad (11) \]

\( N \) is the number of samples, \( r_i \) is the output score, and \( \hat{r}_i \) is the ground-truth score.

Accuracy indicates whether the prediction score is consistent with the ground-truth score in the binary classification task when the dividing line is 5. It is as follows:

\[ ACCURACY = \frac{TP + TN}{P + N} \quad (12) \]

\( TP \) is the true positive samples, \( TN \) is the true negative samples. \( P \) is the positive samples. \( N \) is the negative samples. \( ACCURACY_{|error|\leq1} \) indicates whether the absolute value of the error between the output score and the ground-truth score.
The accuracy score is within 1 point. It can be expressed as follows:

$$\text{ACCURACY}_{\text{error} \leq 1} = \frac{N_{\text{error} \leq 1}}{N}$$  \hspace{1cm} (13)

$N$ is the number of samples, and $N_{\text{error} \leq 1}$ is the number of samples in which the absolute value of the error between the output score and the ground-truth score is within 1 point.

Several experiments were conducted as follows: for training strategies, we record $R$ for the single regression training method, $CR$ for the classification before regression training method, and $PCR$ for classification before regression based on pseudo-labelling training in TABLE III. We verify the goodness of guidance for classification before regression firstly. We use the cropping method in Section III; we can obviously compare $\text{Cropping}(R)$ with $\text{Cropping}(CR)$ to find that if classification before regression will get better result on the testing set. Then we verify that the ECA plays a great role in the aesthetic assessment task. The comparison experiments of $\text{Cropping}(CR)$ and $\text{Cropping+ECA}(CR)$ can verify that the ECA can effectively improve the accuracy of the assessment. Besides, we can verify the training method of pseudo labels can improve the result by $\text{AAB+ECA}(PCR)$. Finally, we can also verify that the meta reweighting network can improve the result by $\text{MRN+}A\text{AB+ECA}(PCR)$. All the ablation experiments results are shown in TABLE III.

Fig. 5 compares the test score results for four main methods. We select some typical samples in each score segment: (a) Anticlockwise: 0.632 Clockwise: 0.627 (b) Forward: 0.533 (c) Backward: 0.547 (c) Look up: 0.518 Look down: 0.529

![Path planning samples. The better path is identified in bold and the best capture in this path is identified by a yellow star.](image)

**TABLE III**

| Method                  | MSE$\downarrow$ | SROCC$\uparrow$ | Acc$\uparrow$ | Acc$_{\text{error} \leq 1}$ $\uparrow$ |
|------------------------|-----------------|-----------------|---------------|--------------------------------------|
| MRN+\text{AAB+ECA}(PCR)| 0.6691          | 0.8471          | 86.53%        | 80.61%                               |
| \text{AAB+ECA}(PCR)    | 0.6888          | 0.8409          | 86.48%        | 79.80%                               |
| \text{AAB+ECA}(CR)     | 0.7575          | 0.8238          | 84.58%        | 78.73%                               |
| \text{Resizing+ECA}(CR)| 0.7684          | 0.8213          | 84.82%        | 78.23%                               |
| \text{Cropping+ECA}(CR)| 0.8213         | 0.8117          | 84.43%        | 76.54%                               |
| \text{Cropping}(CR)    | 1.0325          | 0.7743          | 79.23%        | 69.58%                               |
| \text{Cropping}(R)     | 1.2520          | 0.7359          | 76.39%        | 63.78%                               |

**TABLE IV**

| Method      | MAE$\downarrow$ | SROCC$\uparrow$ | Acc$\uparrow$ |
|-------------|-----------------|-----------------|---------------|
| Reg.Net [4] | 0.1268          | 0.678           | -             |
| RGNet [31]  | 0.1110          | 0.710           | -             |
| Lee et al. [32] | 0.1141      | 0.879           | -             |
| Ours        | 0.1016          | 0.7185          | 80.5%         |

$G$ represents the ground truth; $M$ represents the score of $\text{MAE}(PCR)$; $A$ represents the score of $\text{AE}(PCR)$; $R$ represents the score of $\text{RE}(CR)$; $C$ represents the score of $\text{CE}(CR)$. It can be found that each segment of the dataset has obvious distinction, and $\text{MAE}(PCR)$ method has the best ability for...
Fig. 7. Comparison of scatter plots. 6(a) represents Cropping(CR), 6(b) represents Cropping + ECA(CR), 6(c) represents Resizing + ECA(CR), 6(d) represents AAB + ECA(CR), 6(e) represents AAB + ECA(PCR), and 6(f) represents MRN + AAB + ECA(PCR). The horizontal axis represents the prediction scores, and the vertical axis represents the ground-truth scores.

TABLE V
PREDICTION RESULTS OF DIFFERENT IMAGE PREPROCESSING METHODS

| Image size | Origin | AAB | Resizing | Cropping |
|------------|--------|-----|----------|----------|
| 1000 x 582 | 8.12   | 8.14| 8.21     | 6.78     |
| 1600 x 582 | 8.12   | 7.67| 8.21     | 6.97     |
| 1000 x 1782| 8.12   | 7.80| 8.29     | 7.33     |
| 250 x 146  | 8.12   | 7.83| 8.02     | 7.23     |

We can get the prediction of the distribution with different methods from the scatter plots of Fig.7. Fig.7 shows that the best model can predict the scores between 1 and 9, which is very close to the actual label score interval; the prediction results of the high and low segments are ideal; (f) shows a long and straight spinning cone, which means that (f) has the best distribution for aesthetic regression.

V. CONCLUSION

It is a challenge task to construct a new dataset in image aesthetic field. We construct AMD-CR with reasonable classification regression by mixing and filtering massive datasets; we proposed a new aesthetic evaluation model based on meta reweighting network and binary classification pseudo labels. This model can effectively improve ability of predicting image aesthetic. Moreover, we design AAB structure to pretreat in any input size images. We are also the first team use ECA in aesthetics. This attention can improve ability of extracting features without changing the number of channels. The best SROCC of test set is 0.8471, which is higher than the result of classic deep learning regression model.

Now, we only evaluate the overall score of the images. In the future, we will improve the quality of image aesthetic assessment by studying the interpretability of aesthetic attributes. We also consider how to use emotions, arts, themes, and other high semantic information to help aesthetic evaluation tasks. Moreover, we hope to explore more in the area of aesthetic guidance.

ACKNOWLEDGMENT

The authors would like to thank all the editors and reviewers. They also want to use their model on MindSpore,\(^1\) which

\(^1\)https://www.mindspore.cn/
is a new deep learning computing framework. These problems are left for future work.

REFERENCES
[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Studying aesthetics in photographic images using a computational approach,” in Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer, 2006, pp. 288–301.
[2] A. Brachmann and C. Redies, “Computational and experimental approaches to visual aesthetics,” Front. Comput. Neurosci., vol. 11, pp. 1021, Nov. 2017.
[3] N. Murray, L. Marchesotti, and F. Perronnin, “AVA: A large-scale database for aesthetic visual analysis,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 2408–2415.
[4] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, “Photo aesthetics ranking network with attributes and content adaptation,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 662–679.
[5] K.-Y. Chang, K.-H. Lu, and C.-S. Chen, “Aesthetic critiques generation for photos,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3514–3523.
[6] W. Luo, X. Wang, and X. Tang, “Content-based photo quality assessment,” in Proc. Int. Conf. Comput. Vis., Nov. 2011, pp. 2206–2213.
[7] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, “ECA-Net: Efficient channel attention for deep convolutional neural networks,” in Proc. CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 11531–11539.
[8] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. 36th Int. Conf. Mach. Learn., 2019, pp. 6105–6114.
[9] Q. Kuang, X. Jin, Q. Zhao, and B. Zhou, “Deep multimodality learning for UAV video aesthetic quality assessment,” IEEE Trans. Multimedia, vol. 22, no. 10, pp. 2623–2634, Oct. 2020.
[10] R. W. Sperry, M. S. Gazzaniga, and J. E. Bogen, “Interhemispheric relationships: The neocortical commissures; syndromes of hemisphere disconnection,” in Handbook of Clinical Neurology, no. 4. Amsterdam, The Netherlands: North-Holland, 1969, pp. 273–290.
[11] H. Tong, M. Li, H.-J. Zhang, J. He, and C. Zhang, “Classification of digital photos taken by photographers or home users,” in Proc. Pacific-Rim Conf. Multimedia, Berlin, Germany: Springer, 2004, pp. 198–205.
[12] Y. Luo and X. Tang, “Photo and video quality evaluation: Focusing on the subject,” in Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer, 2008, pp. 386–399.
[13] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, “Aesthetic quality classification of photographs based on color harmony,” in Proc. CVPR, Jun. 2011, pp. 33–40.
[14] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, “Assessing the aesthetic quality of photographs using generic image descriptors,” in Proc. Int. Conf. Comput. Vis., Nov. 2011, pp. 1784–1791.
[15] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, “RAPID: Rating pictorial aesthetics using deep learning,” in Proc. 22nd ACM Int. Conf. Multimedia, 2014, pp. 457–466.
[16] Y. Kao, C. Wang, and K. Huang, “Visual aesthetic quality assessment with a regression model,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2015, pp. 1583–1587.
[17] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, “Deep multi-patch aggregation network for image style, aesthetics, and quality estimation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 990–998.
[18] O. Wu, W. Hu, and J. Gao, “Learning to predict the perceived visual quality of photos,” in Proc. Int. Conf. Comput. Vis., Nov. 2011, pp. 225–232.
[19] X. Jin et al., “Predicting aesthetic radar map using a hierarchical multi-task network,” in Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV). Cham, Switzerland: Springer, 2018, pp. 41–50.
[20] X. Jin et al., “Incremental learning of multi-tasking networks for aesthetic radar map prediction,” IEEE Access, vol. 7, pp. 183647–183655, 2019.
[21] G. Fini, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in Proc. 34th Int. Conf. Mach. Learn., 2017, pp. 1126–1135.
[22] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, “Meta-learning with memory-augmented neural networks,” in Proc. 33rd Int. Conf. Mach. Learn., 2016, pp. 1842–1850.
[23] J. Andrychowicz et al., “Learning to learn by gradient descent by gradient descent,” 2016, arXiv:1606.04474.
[24] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, “Matching networks for one shot learning,” 2016, arXiv:1606.04080.
[25] S. Ravi and H. Larochelle, “Optimization as a model for few-shot learning,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2017. [Online]. Available: https://openreview.net/forum?id=rJY0-Kc1l
[26] N. Schweighofer and D. Doya, “Meta-learning in reinforcement learning,” Neural Netw., vol. 16, no. 1, pp. 5–9, Jan. 2003.
[27] J. Shu et al., “Meta-weight-net: Learning an explicit mapping for sample weighting,” 2019, arXiv:1902.07379.
[28] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.
[29] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015, arXiv:1503.02531.
[30] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, “Perceptual quality assessment of smartphone photography,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 3677–3686.
[31] D. Liu, R. Puri, N. Kamath, and S. Bhattacharyya, “Composition-aware image aesthetics assessment,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2020, pp. 3569–3578.
[32] J.-T. Lee and C.-S. Kim, “Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., Oct. 2019, pp. 1191–1200.

Xin Jin was born in Anhui, China. He received the Ph.D. degree in computer science from Beijing University of Technology, China, in 2013. He was a Visiting Student at the Lotus Hill Institute, China, and a Visiting Scholar at Tsinghua University, Beijing, China. He is currently an Associate Professor with the Department of Cyber Security, Beijing Electronic and Science Technology Institute, and a Visiting Scholar at the Beijing Institute for General Artificial Intelligence (BIGAI). His research interests include computational aesthetics, computer vision, and artificial intelligence security. He served as a Program Committee Member and the Session Chair for multiple conferences, including AAAI 2020–2022, IJCAI 2020, ACM MM 2021, ISAIR 2017–2022, and ICVRD 2017–2020. He has received the Best Paper Award from ROSENET 2018/2019.

Hao Lou was born in Zhejiang, China. He received the bachelor’s degree from Northwest Agriculture and Forestry University, China. He is currently pursuing the master’s degree with the Beijing Electronic Science and Technology Institute, China. His research interests include visual media aesthetics.

Heng Huang was born in Fujian, China. He received the bachelor’s degree from Shandong University, China. He is currently pursuing the master’s degree with the Beijing Electronic Science and Technology Institute, China. His research interests include visual media aesthetics.
Xinning Li was born in Shandong, China. He received the bachelor’s degree from the Beijing Electronic Science and Technology Institute, China, where he is currently pursuing the master’s degree. His research interests include visual media aesthetics.

Xiaodong Li was born in Henan, China. He received the Ph.D. degree from Northwestern Polytechnic University, China. He is currently an Associate Professor with the Beijing Electronic Science and Technology Institute, China. His researches are focused on information security and visual media security.

Shuai Cui was born in Beijing, China. She received the Bachelor of Arts and Sciences degree in philosophy and mathematics from the University of California, Davis, CA, USA, in 2021. She is currently pursuing the master’s degree. Her research interests include logic, metaphysics, epistemology, and artificial intelligence.

Xiaokun Zhang was born in Gansu, China. He is currently a Professor with the Department of Cyber Security, Beijing Electronic Science and Technology Institute, China. His research interests include computer science and technology.

Xiqiao Li was born in Shanxi, China. He received the master’s degree from the Beijing Electronic Science and Technology Institute, Beijing, China. His research interests include visual media aesthetics.