Aspect-Ratio-Preserving Multi-Patch Image Aesthetics Score Prediction

Lijie Wang Xueting Wang Toshihiko Yamasaki Kiyoharu Aizawa
The University of Tokyo, Japan
{wang, xt_wang, yamasaki, aizawa}@hal.t.u-tokyo.ac.jp

Abstract

Owing to the spread of social networking services (SNS), there is an increasing demand for automatically selecting, editing or generating impressive images, which raises the importance of evaluating image aesthetics. We propose the first multi-patch method for image aesthetic score prediction with the original image aspect ratios being preserved. Our method just uses images for training and does not require external information both in training as well as prediction. In an experiment using the large-scale AVA dataset containing 250,000 images, our approach outperforms other existing methods in image aesthetic score prediction, especially reducing mean squared error (MSE) of predicted aesthetic scores by 0.061 (18%) and improving the linear correlation coefficient (LCC) by 0.056 (8.9%). Noticeably, the decrease in mean absolute error (MAE) by our method for images with an unbalanced aspect ratio is at most 7.9 times larger than the decrease in MAE for images with a typical digital camera aspect ratio. This result indicates that our multi-patch method expands the range of aspect ratios with which aesthetics scores of images can be predicted accurately.

1. Introduction

Owing to the widespread popularity of social networking services (SNS), there is an increasing demand for uploading attractive images to SNS. However, as many users do not have skills to select, edit and generate aesthetic images, there has been a substantial request for an automatic process that lets one obtain aesthetic images. To realize this, a key element is to accurately assess the aesthetics of images.

Nevertheless, aesthetic assessment is challenging as aesthetics highly depend on human subjectivity. To assess this obscure sensitivity, it is necessary to extract features from the entire image and combine them appropriately.

Aesthetic assessment has been studied by many researchers, and various feature extracting methods have been attempted. Among the initial attempts [4,5,13,17,21], handcrafted features about such as object composition and color harmony are designed and used. Following them, according to the success of convolutional neural networks (CNNs) on object recognition tasks, many studies [2, 7, 11, 14, 15, 18, 19, 23, 29–31, 36] have adopted CNNs as feature extractors. Other ideas of CNN architectures such as Siamese-like network [2, 31] and triplet loss [29] have also been applied to aesthetic assessment [13, 14, 30]. We also use a CNN as the image feature extractor.

Except for features from images, extra information is also included to improve predicting accuracy: scene or style annotations in a dataset [7,11,15,18,19,23], multimodal text comments [36], object tags [27], and saliency maps [22]. While those extra characteristics improve aesthetic assessment performance, they lead to the high cost of creating new datasets and limitations of applying models to other tasks as specific extra information is required by the specific model at the training phase, or sometimes at the evaluation phase. In this study, we focus on a fundamental and versatile approach to effective image feature extraction for aesthetics assessment. Therefore, we only used images to predict aesthetics scores either during training or evaluation.

There are three kinds of tasks studied for aesthetics assessment: positive/negative binary classification task [20, 23], aesthetics rating distribution prediction task [3, 10], and aesthetics score prediction task [15, 34]. In this paper, we conduct aesthetics score prediction. Aesthetics score prediction is useful for quantitative evaluation applications such as recommendation systems, in contrast to aesthetic binary classification. An aesthetics score of the image is calculated as the mean of its aesthetics rating distribution, which is labeled by humans and usually provided in a dataset. The samples images, normalized rating histograms, and aesthetic scores of the AVA dataset [24], a large-scale aesthetics dataset, are shown in Fig. 1. Aesthetics score prediction has been conducted by Kao *et al*. [12], Jin *et al*. [9], Roy *et al*. [27], and Talebi *et al*. [34]. However, those methods all rescale images to square images regardless of their original aspect ratios, including the most outstanding method called NIMA proposed by Talebi *et al*. [34]. The lack of aspect ratio information can affect the prediction of aesthetics scores, especially for those images having un-

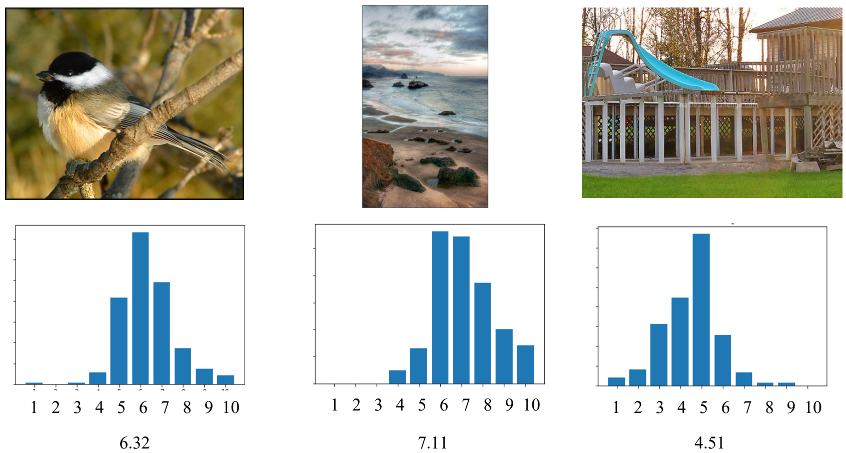

Figure 1: Sample images (top), normalized rating histograms (middle), and means of the rating histograms calculated as aesthetics scores (bottom) from the AVA dataset. Each column shows a pair of them.

usual aspect ratios. Furthermore, it can easily cause contradictions with human aspect-ratio-dependent aesthetics.

To resolve this problem, we propose an aspect-ratio-preserving multi-patch learning for aesthetics score prediction. We crop several patches from an input image, predict normalized aesthetics rating distributions for each patch, and calculate the aesthetics score by aggregating these distributions. In the training, we use the multi-patch earth mover's distance (EMD) as a part of the loss function. Using the AVA dataset [24], which has more than 250,000 images, our experimental results demonstrate that aspect-ratio-preserving multi-patch learning improves the performance of aesthetics score prediction. Our method reduces the mean squared error (MSE) by 0.061 (18%) compared to a simple CNN-based method [9], and improves the linear correlation coefficient (LCC) of aesthetics scores by 0.056 (8.9%) and the Spearman's rank correlation coefficient (SRCC) by 0.074 (12%) compared to the existing method NIMA [34]. Furthermore, using our method, the mean absolute error (MAE) of prediction for images with unusual aspect ratios is improved significantly.

In summary, our main contributions are as follows:

- We are the first to propose aspect-ratio-preserving multi-patch learning approach for predicting aesthetics scores, in order to reflect the original aspect ratio information to prediction.
- Experimental results demonstrate that our method reduces the MSE by 0.061 (18%), increases the LCC of aesthetics scores by 0.056 (8.9%), and increases the SRCC by 0.074 (12%) compared to the existing methods. Especially, our method demonstrated the significant improvement for images with unusual aspect ratios.
- Our versatile method uses images and aesthetic ratings without extra information to achieve high performance of predicting the aesthetic scores, for maintaining applicability to other datasets and other tasks.

2. Related works

Aesthetic assessment can be broadly categorized into three tasks: high/low aesthetic binary classification, aesthetics rating distribution prediction, and prediction of the mean of the rating distribution. The mean of the rating distribution is usually called as "aesthetics score". High/low aesthetics binary classification is tackled by many studies [11, 15, 17–23, 30, 32, 36], but there have only been a few studies on rating distribution prediction [3, 6, 10, 35] and aesthetics score prediction [9, 12, 27, 34]. From here, we explain previous works related to our task: aesthetics score prediction.

Aesthetics score prediction Among aesthetics score prediction, to the best of our knowledge, the first attempt to predict aesthetics score was made by Kao *et al*. [12] using a regression network. This network comprises five convolution layers and four fully connected (fc) layers, and directly predicts the aesthetics score of the image. Jin *et al*. [9] trained network by adding large weights to images with rare aspect ratios in the dataset. Roy *et al*. [27] also used extra object tags to predict aesthetics scores. In contrast, instead of directly regressing aesthetic score as these methods, Talebi *et al*. [34] proposed NIMA, an approach that calculates aesthetics scores from predicted aesthetics rating distributions. NIMA has two outstanding novelties. The first is that NIMA uses rating distributions to use more information about ratings compared to direct aesthetics score regression.

Table 1: Comparison of functions among previous aesthetics assessment works and our method.

	NIMA [34]	$\mathrm{MP_{ada}}$ [32]	*ours*
score prediction	✓		✓
aspect ratio keeping		✓	✓

The second is that NIMA adopted the earth mover's distance (EMD) [8, 16] for training NIMA parameters. EMD is a distribution distance function considering inter-class relationships. Therefore, the model can learn the global characteristics of distributions, without sticking to fitting local values of distributions elaborately.

However, due to the restriction of the CNN, all images are rescaled to square images to feed into the network regardless of their aspect ratios. By this transformation, images lose their aspect ratio information. It can affect the prediction of aesthetics scores, especially those images having unusual aspect ratios. Furthermore, this contradicts the fact that the NIMA network predicts the same aesthetics score to the original image and the rescaled image, whereas humans can easily find a decrease in aesthetics for the rescaled image.

Multi-patch learning To resolve this problem, aspect-ratio-preserving multi-patch learning is a promising approach. For the high/low aesthetic binary classification task, some multi-patch methods have been proposed [20, 22, 23, 32]. Among them, Sheng *et al*. [32] proposed a weighted multi-patch aggregation system for the output of each patch with the original aspect ratio, which is the latest and highly effective method. Using this system, the network is trained strongly from wrongly predicted patches. In this connection, spatial pyramid pooling (SPP) is another possible solution for maintaining aspect ratio. However, as Lu *et al*. [20] demonstrated that SPP did not make significant contributions to aesthetics assessment, we do not adopt SPP.

However, multi-patch learning has been only applied to aesthetic binary classification. We applied the aspect-ratio-preserving multi-patch learning to predict aesthetics scores by predicting normalized aesthetics rating distributions. The brief comparison of functions among NIMA [34], $\mathrm{MP_{ada}}$ proposed by Sheng *et al*., and our methods is shown in Table 1.

3. Proposed Method

In this section, we introduce our training and prediction system for assessing aesthetics scores. We first describe the architecture of our method and continue to explain the proposed loss functions in detail.

3.1. Multi-patch training/evaluation flow

The structure of our proposed multi-patch method is shown in Fig. 2. In the training phase, a fixed number of square patches with the original aspect ratio are first cropped at random from an input image. By extracting patches with original aspect ratios, the model can learn image feature extraction with the same aspect ratio as humans see. Therefore, it is considered to be easier to learn the human subjectivity of aesthetics. Furthermore, the model is expected to be trained effectively, without disturbance made by the uniform square reshape in spite of original aspect ratios which happens in related works such as NIMA [34]. The extracted aspect-ratio-preserved patches are fed into the model and distributions of aesthetics ratings are predicted for each patch. The sum of each distribution is normalized to 1 by calculating a softmax function over the output of the last fc layer. EMD (Eq. (1)) is calculated for each rating distribution, and the loss value is computed by aggregating EMDs from each patch using one of the loss functions described in Section 3.2. Model parameters are updated by backpropagation using this loss value, and these updates are repeated for several epochs using different cropped patches. In the evaluation phase, rating distributions predicted with patches from each image are averaged simply, and an aesthetic score is calculated as the mean of the simple averaged rating distribution.

3.2. Loss function

Earth mover's distance (EMD) As a distance function between rating distributions, we use earth mover's distance (EMD) just like NIMA [34]. EMD is a distance function between two distributions. Unlike cosine similarity or KL divergence, EMD can consider distance among classes. Therefore, the model can learn the global properties of rating distributions, without being bound to fit local value of each class elaborately. An r-norm EMD distance is defined as the minimum cost of transporting values from one distribution to the another, where the distance between the i-th class s_t and the j-th class s_j is calculated as $\|s_i - s_j\|_r$, on the assumption that two distributions have the same classes in the same order.

For N-class aesthetics ratings, if the value of the i-th rating class s_i is i where $1 \le i \le N$, the distance between the i-th rating class s_i and the j-th class s_j is calculated as $|i-j|^r$. In that case, as shown by Levina *et al*. [16], r-norm EMD between two normalized aesthetics rating distributions is calculated as follows:

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

where $\mathrm{CDF}_{\mathbf{p}/\hat{\mathbf{p}}}(k)$ denotes the cumulative distribution function of the ground truth rating distribution $\mathbf{p}$ and

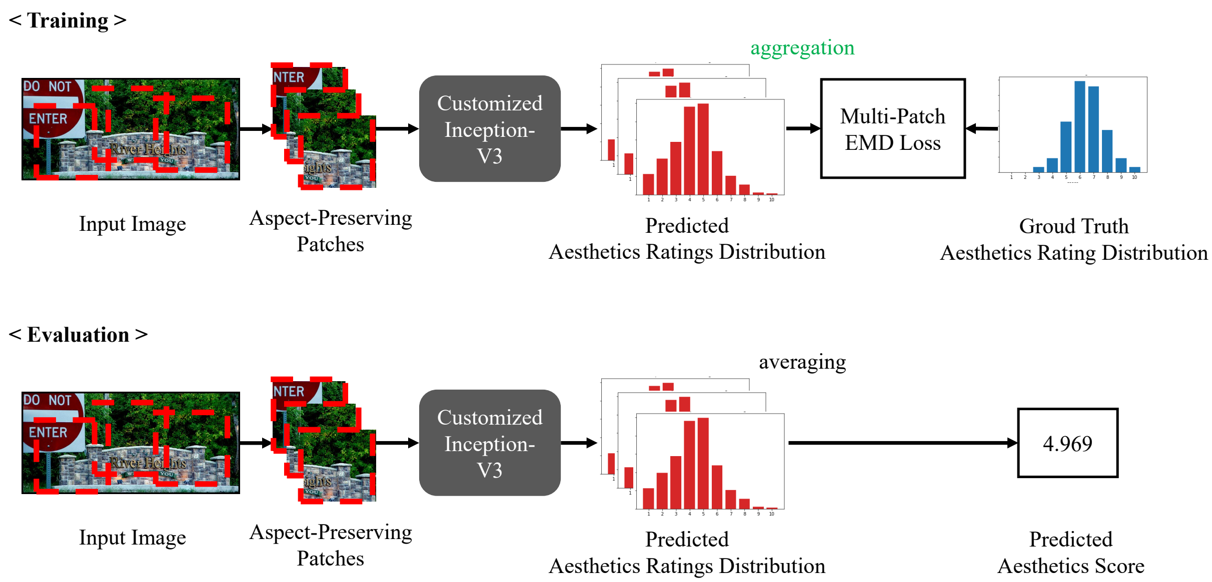

Figure 2: Multi-patch training/evaluation structure of our method.

the predicted rating distribution $\hat{\mathbf{p}}$, which are defined as $\sum_{k=1}^{N} \mathbf{p}$ and $\sum_{k=1}^{N} \hat{\mathbf{p}}$, respectively. We specified r as 2 as well as NIMA.

Multi-patch aggregation We refer to the method proposed by Sheng *et al*. for multi-patch aggregation, which outperforms the other previous works at the high/low aesthetic binary classification task. Compared with the loss function used by Sheng *et al*., we adopt logarithmic 2-norm EMD ($\mathrm{EMD}^{(2)}$, hereinafter, this is just referred to as EMD) to calculate the loss of predicted rating distributions in place of log probability [32] for the binary classification. We use logarithmic EMD instead of mere EMD, expecting a logarithmic function to accelerate training. We propose the two loss functions $\mathrm{MPEMD}_{\mathrm{avg}}$ and $\mathrm{MPEMD}_{\mathrm{ada}}$. $\mathrm{MPEMD}_{\mathrm{avg}}$ simply averages the logarithmic EMDs of plural patches. $\mathrm{MPEMD}_{\mathrm{ada}}$ calculates a weighted mean of the logarithmic EMD to aggregate patches adaptively. These loss functions are defined as follows:

$$\mathrm{MPEMD}_{\mathrm{avg}} = -\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \log \left(\mathrm{EMD_c} \right), \tag{2}$$

$$\mathrm{MPEMD}_{\mathrm{ada}} = -\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \omega_\beta \cdot \log \left(\mathrm{EMD_c} \right), \tag{3}$$

where $\mathcal{P}$ is a set of cropped square patches, p denotes each patch, and $\mathrm{EMD_c}$ is a variable converted from original EMD to represent a kind of certainty of predicted rating distributions. The purpose of training is to minimize EMD which is equivalent to maximizing $\mathrm{EMD_c}$. $\mathrm{EMD_c}$ is defined as follows:

$$\mathrm{EMD_c} = \begin{cases} \epsilon, & (1 - k \cdot \mathrm{EMD} < \epsilon) \\ 1 - k \cdot \mathrm{EMD}, & (\epsilon \leq 1 - k \cdot \mathrm{EMD}) \end{cases} \tag{4}$$

where ϵ is an appropriately small positive constant and k is an expansion coefficient. $\mathrm{EMD_c}$ takes values close to 1 when EMD is low and takes values near 0 when EMD is high. The value of $\mathrm{EMD_c}$ is restricted to $[\epsilon,\ 1]$. The hyper-parameter k is used to adjust the sensitivity of the converted certainty variable $\mathrm{EMD_c}$ to EMD. As the increase of k, the variation of EMD causes a larger change of $\mathrm{EMD_c}$.

ω_β is introduced as the weight of patches and defined as:

$$\omega_\beta = 1 - {\mathrm{EMD_c}}^{\beta}. \tag{5}$$

ω_β is high when the certainty variable $\mathrm{EMD_c}$ is low, and vice versa. The value of ω_β ranges from 0 to 1. The hyper-parameter β ($\beta > 0$) determines the range of $\mathrm{EMD_c}$ with which patches are trained strongly. Fig. 3 shows how the patch weight ω_β varies with the certainty variable $\mathrm{EMD_c}$ for each β. For example, as shown in Fig. 3, if β is large, patches with large $\mathrm{EMD_c}$ are even weighted heavily. This means patches with small $\mathrm{EMD_c}$ are also strongly trained.

The effect of k and β is dependent on each other; thus k and β should be optimized together.

4. Experiment

In this section, we first describe the dataset used in our experiment. Then, we explain training configurations for three experiments: pre-training with NIMA, and training using $\mathrm{MPEMD}_{\mathrm{avg}}$ and $\mathrm{MPEMD}_{\mathrm{ada}}$. Finally, we present

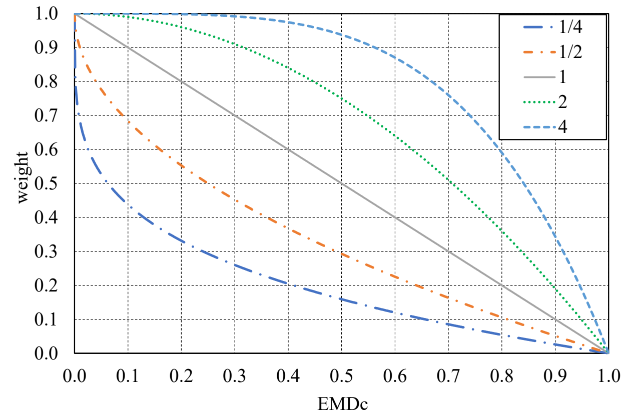

Figure 3: Relationship between the patch weight ω_β and the certainty variable $\mathrm{EMD_c}$, with respect to each β.

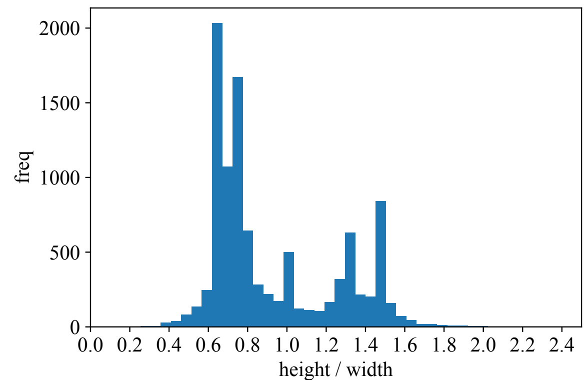

Figure 4: Histogram of aspect ratios (height/width) of images in sampled AVA dataset.

the results of our experiments and comparisons between our study and previous works.

4.1. Dataset

We trained and evaluated our proposed models using the AVA dataset [24]. The AVA dataset comprises 250,000 images collected from the online photography community website `www.dpchallenge.com`. Each image is associated with 10 stages of ratings, ranging from 1 to 10. The number of raters assigned to each image ranges from 78 to 649, and the average value is 210. Samples of the AVA dataset, including images, normalized rating histograms, and means of the rating histograms, called as aesthetic scores, are shown in Fig. 1. Except for ratings, some images have additional attributes such as semantic and photographic style information, which were neither used for training nor testing in our experiment.

Fig. 4 shows the histogram of aspect ratios (height/width) of the images in the AVA dataset. As shown in Fig. 4, most of the images have aspect ratios from 0.6 to 0.8. Especially, there are two peaks within the ranges of 0.62 to 0.67 and 0.72 to 0.77. This concentration can be explained by the fact that normal digital cameras are configured to take photos with the ratio of the image height to the image width as 2:3 (the aspect ratio is 0.66) or 3:4 (the aspect ratio is 0.75). In other words, the AVA dataset contains relatively a small number of images with their aspect ratios not falling within the range 0.6 to 0.8, which means those aspect ratios have less training images.

We used the AVA dataset [24] for both training and evaluation. The AVA dataset we used contains 255,494 pairs of an image and a rating histogram. In the same way as previous multi-patch works [20, 23, 32], we used 92 % of the entire dataset for training. Additionally, half of the remaining dataset (4% of the entire dataset) was used for test and the other half (4% of the entire dataset) was assigned for validation. Therefore, 235,054 images were used for training, 10,220 images were used for validation, and the other 10,220 images were used in the test dataset. It should be noted that some other previous works used different numbers of images for training/validation/test datasets. For example, Kao *et al.* [12], Jin *et al.* [9], Roy *et al.* [27] used about 250,000 images for the training and 5,000 images for the test, and Talebi *et al.* [34] used about 204,000 images for the training of NIMA and 51,000 images for the test. The reason we chose this partition (92:4:4) is that 5,000 test images were not enough for the analysis about aspect ratios described in Section 5 and 51,000 images are too many for the test. For a fair comparison, we also show the result of reimplemented NIMA trained with 92% of the entire AVA dataset in Section 5.

4.2. Training

Pre-training was conducted using the same architecture as NIMA and the AVA training set. We use a customized Inception-V3 [33] with the last fully connected (fc) layer replaced by a randomly initialized fc layer with 10 output channels, as the CNN image feature extractor. All layers apart from the last new fc layer were initialized by the parameters pre-trained on the ImageNet dataset [28]. All images from the training set are resized to 342×342, after which 299×299 random cropping and random horizontal flipping were applied as data augmentations. We set the learning rate to 10^{-3} instead of 3×10^{-7} and 3×10^{-6}, reported by Talebi *et al.* [34], because the model could not be trained adequately in our environment using those learning rates. For the other training settings, we used a momentum SGD optimizer with the momentum of 0.9, and let learning rate decay by a factor of 0.95 after every 10 epochs. We trained the model for 100 epochs.

Following this, the aspect-ratio-preserving multi-

Table 2: Comparison of the aesthetics score prediction performance of our methods and those of previous works. The first eight rows present the results of previous works and the bottom three rows indicate the results of our experiments. For each metric, the best value is shown in bold.

Models	LCC ↑	SRCC ↑	MSE ↓	acc [%] ↑	EMD ↓
GIST linear-SVR [12]	-	-	0.0522	-	-
GIST RBF-SVR [12]	-	-	0.5307	-	-
BOV-SIFT linear-SVR [12]	-	-	0.5401	-	-
BOV-SIFT RBF-SVR [12]	-	-	0.5513	-	-
Kao *et al.* [12]	-	-	0.4510	-	-
Jin *et al.* [9]	-	-	0.3373	-	-
Roy *et al.* [27]	-	-	0.3562	-	-
NIMA (Inception-V2) rept. 2018 [34]	0.636	0.612	-	**81.51**	**0.050**
NIMA (our impl. using Inception-V3)	0.6914	0.6802	0.2830	79.88	0.066
$\mathrm{MPEMD}_{\mathrm{avg}}$ (ours)	0.6900	0.6854	0.2788	79.08	0.065
$\mathrm{MPEMD}_{\mathrm{ada}}$ (ours)	**0.6923**	**0.6868**	**0.2764**	79.38	0.066

patch training was conducted using the loss functions $\mathrm{MPEMD}_{\mathrm{avg}}$ and $\mathrm{MPEMD}_{\mathrm{ada}}$. The same customized Inception-V3 was used as the CNN image extractor and all layers were initialized by pre-trained NIMA parameters. The reason we used these parameters was that the model could efficiently obtain the feature extraction ability on both the global composition and local fine-grained features. This was expected to shorten the training time for multi-patch learning. Input patches were extracted in the following manner: first, we resized the shorter edge of every image in the dataset to 342 pixels while keeping its aspect ratio; then extracted 8 299×299 crops from each rescaled image. The learning rate was set to 10^{-3}. For the loss function hyperparameters, we set k to 1.2 and β in $\mathrm{MPEMD}_{\mathrm{ada}}$ to 0.4, based on the hyperparameter tuning using Tree-structured Parzen Estimator (TPE) [1] implemented by Optuna [26]. For the other learning settings, we used a momentum SGD optimizer with momentum of 0.9 and the weight decay rate of 10^{-4} and let the learning rate decay by a factor of 0.7 after every 10 epochs. We trained the model for 50 epochs.

All models were implemented using PyTorch v.0.4.1 [25].

5. Results

First, we demonstrate the overall performance of our methods using several metrics and via comparison with previous works. Following that, we demonstrate MAE improvement for each aspect ratio by aspect-ratio-preserving multi-patch learning.

Overall performance We used linear correlation coefficient (LCC), Spearman's rank correlation coefficient (SRCC), and mean squared error (MSE) for evaluating the aesthetics score prediction performance of our three experiments and those of previous works. Additionally, we calculated accuracy (acc) of aesthetics binary classification and average EMD for comparison with NIMA [34]. For binary classification, images with aesthetics scores less than or equal to 5 are labeled as negative and the rest are labeled as positive. Nonetheless, it should be kept in mind that the main purpose of this study is aesthetics score prediction.

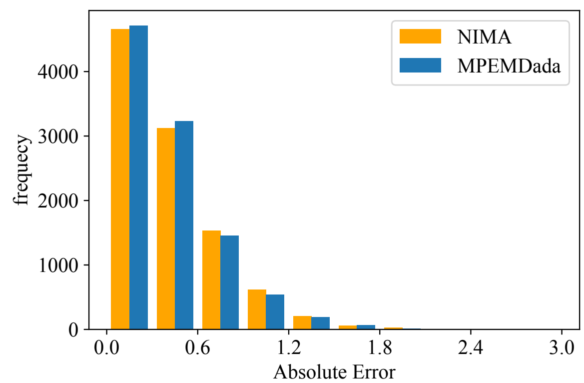

Figure 5: Comparison of histograms of absolute errors (AEs) of aesthetics scores predicted for the test dataset between the $\mathrm{MPEMD}_{\mathrm{ada}}$ model and the pre-trained NIMA.

The results are shown in Table 2. The model trained with the loss function $\mathrm{MPEMD}_{\mathrm{ada}}$ outperforms previous works for all metrics evaluated for aesthetics score prediction. Compared with the previous best result corresponding to each metric, LCC shows an improvement of 0.056 (8.9%) and SRCC shows an improvement of 0.074 (12%) compared to the performance of NIMA reported by Talebi *et al.* [34]; and MSE shows an improvement of 0.061 (18%) compared to the performance reported by Jin *et al.* [9].

Among our experiments, the model trained with the loss

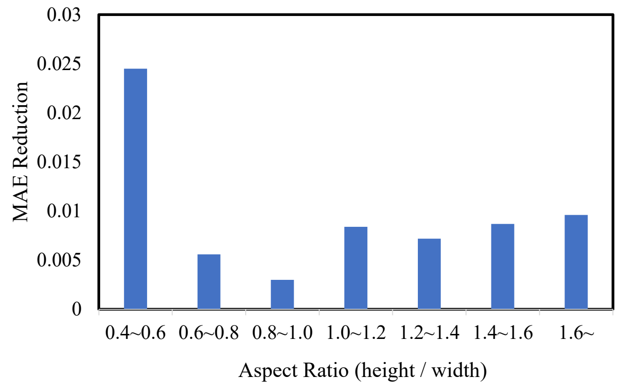

Figure 6: Average aesthetics score MAE reduction of each aspect ratio by the model trained with $\mathrm{MPEMD}_{\mathrm{ada}}$ compared to the pre-trained NIMA model.

function $\mathrm{MPEMD}_{\mathrm{ada}}$ outperforms the other two our experiments. Comparing the model trained with $\mathrm{MPEMD}_{\mathrm{ada}}$ and the pre-trained NIMA model, the SRCC and MSE of the $\mathrm{MPEMD}_{\mathrm{avg}}$ model is better than that of the pre-trained NIMA model, while the LCC is worse. Overall, considering an MSE decrease of 0.0042 (1.5%) which is the most significant among that of the three metrics, it can be argued that the model trained with $\mathrm{MPEMD}_{\mathrm{avg}}$ outperforms the pre-trained NIMA model. This indicates that the aspect-ratio-preserving multi-patch training is efficient for aesthetics score prediction even with the simple average aggregation of plural patches. In addition, the fact that the model trained with $\mathrm{MPEMD}_{\mathrm{ada}}$ outperforms the model trained with $\mathrm{MPEMD}_{\mathrm{avg}}$ for all metrics evaluated demonstrates that weighted multi-patch aggregation also improves aesthetics score prediction performance. However, no improvement is shown in the accuracy of aesthetic binary classification and the optimization of EMD. The performance of NIMA reported by Sheng *et al*. is superior to the performance of our methods.

As a reference, the histogram of absolute errors (AEs) predicted by the $\mathrm{MPEMD}_{\mathrm{ada}}$ model and the pre-trained NIMA model for the test dataset is shown in Fig. 5. Fig. 5 demonstrates that predicted aesthetics scores contain their AEs within 0.3 for approximately 45 % of test images and within 0.66 for more than 75% of test images. Furthermore, the number of images with small AEs predicted by the $\mathrm{MPEMD}_{\mathrm{ada}}$ is larger than the number of those predicted by the pre-trained NIMA model, and the number of images with middle AEs predicted by the $\mathrm{MPEMD}_{\mathrm{ada}}$ is smaller than the number of those predicted by the pre-trained NIMA model. Therefore, it can be also found in Fig. 5 that $\mathrm{MPEMD}_{\mathrm{ada}}$ decrease error of aesthetics score prediction.

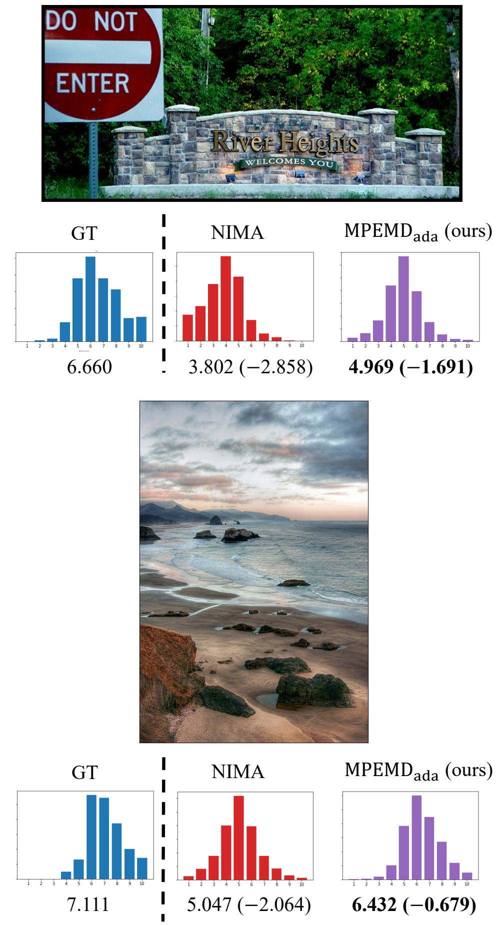

Figure 7: Examples of prediction improved by the $\mathrm{MPEMD}_{\mathrm{ada}}$ model compared to the NIMA model. Numbers under distribution denote aesthetic scores and the number inside each bracket is the difference between the prediction and the ground truth.

Dependence of MAE improvement on image aspect ratios We also investigated the MAE improvement corresponding to each aspect ratio by the model trained with $\mathrm{MPEMD}_{\mathrm{ada}}$ from the MAE of the pre-trained NIMA model. The results are shown in Fig. 6. Fig. 6 demonstrates that aesthetics score prediction improves significantly for images with aspect ratios (height/width) lower than 0.6 or higher than 1.0. For example, the decrease in MAE for images with aspect ratios within the range 0.4 to 0.6 is 7.9 times larger than the decrease for images with aspect ratios within the range 0.8 to 1.0. As described in Section 4.1, those aspect ratios are unusual in the AVA dataset. This can be ascribed to the ability of the multi-patch trained model to use the information of an original aspect ratio of the image, in contrast to the NIMA model which ignores aspect ratio information. Because NIMA does not use aspect ratio

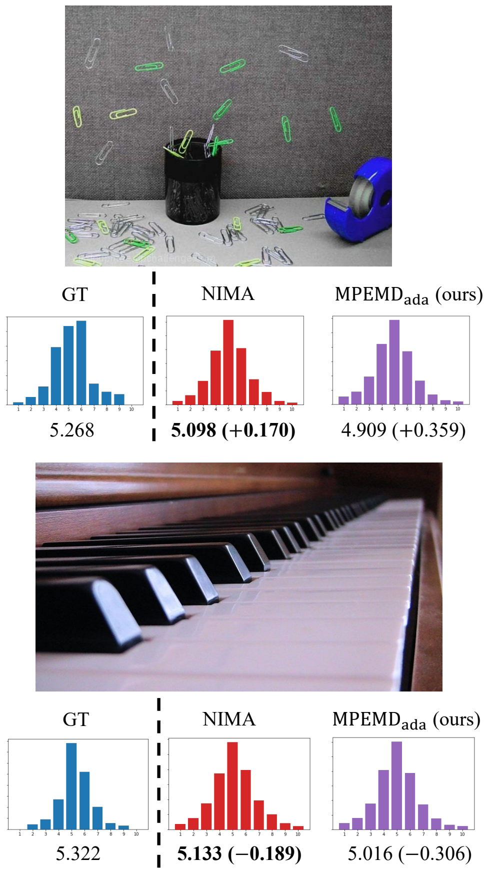

Figure 8: Examples of prediction deterioration by the $\mathrm{MPEMD}_{\mathrm{ada}}$ model compared to the NIMA model. Numbers under distribution denote aesthetic scores, and a number inside each bracket is the difference between the prediction and the ground truth.

information, it tends to fit for images with common aspect ratios and not trained enough to those with unusual aspect ratios. Our method resolves this problem by using multi-patch training with aspect ratios preservation. Therefore, for a variety of aspect ratios, aesthetics scores could be predicted accurately by our method. Furthermore, the result indicates that aesthetics scores of images which have rare aspect ratios in the training dataset also can be predicted accurately, which have been hard to be predicted by existing methods.

Examples of improved prediction by the $\mathrm{MPEMD}_{\mathrm{ada}}$ model are shown in Fig. 7 and examples of deteriorated prediction are shown in Fig. 8. Particularly, the aesthetics score predictions of the first image in Fig. 7, which is quite lengthy horizontally, and the third image, which is quite long vertically, are significantly improved with the use of the $\mathrm{MPEMD}_{\mathrm{ada}}$ model.

Discussion As described above, our method using aspect-ratio-preserving multi-patch learning and prediction outperforms previous works in aesthetics score prediction performance. Furthermore, our method improves aesthetics score prediction for images with unusual aspect ratios, and it leads to the expansion of the range of aspect ratios with which aesthetics scores can be predicted accurately. However, errors still remain. Some of which are inevitable as human aesthetics are subjective, but we believe the difference in the shape of distribution between the ground truth and the prediction is worth mentioning. Peculiarly, distributions with a peak extending over several ratings, such as the first image in Fig. 8 and the second successful image in Fig. 7, are not well predicted. This point may be addressed by modifying the last activation function, which may improve the aesthetics score prediction performance due to its enhanced ability to generate rating distributions.

Additionally, our methods do not work well for aesthetic binary classification and EMD optimization. The reason for the low performance of binary classification is considered to be the prediction bias around the classification threshold. However, as a slight prediction bias near the classification threshold can largely affect classification accuracy, this result does not conflict with the success of aesthetics score prediction. Besides, the failure in optimizing EMD is also not incompatible with the successful aesthetics score prediction because optimizing EMD is not equal to optimizing score prediction.

6. Conclusion

We proposed methods of aspect-ratio-preserving multi-patch training and prediction to predict the mean of aesthetics rating, which is termed aesthetics score. Using our methods, we were able to reflect the aspect ratio information to the model. From experiments using the AVA dataset, our methods could outperform previous works in all metrics related to aesthetics score prediction performance. In particular, the model trained with our multi-patch weighted loss named $\mathrm{MPEMD}_{\mathrm{ada}}$ reduced the MSE by 0.061 (18%) compared to the best MSE reported by previous works. Especially, our method improves prediction performance for images with unusual aspect ratios. This result indicates that our method enables the model to predict aesthetics scores accurately for a wide range of aspect ratios. Our methods could also be easily applied to other datasets or other tasks, as we do not use any external information both in training and prediction. Generalization of our proposed patch-based method is considered to be the next development.

Acknowledgments. This research is partially supported by JST-CREST (JP-MJCR1686) and JSPS KAKENHI (JP-18H03339).

References

[1] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *Proc. of Neural Information Processing Systems (NIPS)*, pages 2546–2554, 2011.

[2] S. Chopra, R. Hadsell, and Y. LeCun. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 539–546, 2005.

[3] C. Cui, H. Fang, X. Deng, X. Nie, H. Dai, and Y. Yin. Distribution-oriented Aesthetics Assessment for Image Search. In *Proc. of ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 1013–1016, 2017.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying Aesthetics in Photographic Images Using a Computational Approach. In *Proc. of European Conf. on Computer Vision (ECCV)*, volume 3953 LNCS, pages 288–301, 2006.

[5] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 1657–1664. IEEE, 2011.

[6] H. Fang, C. Cui, X. Deng, X. Nie, M. Jian, and Y. Yin. Image Aesthetic Distribution Prediction with Fully Convolutional Network. In *Proc. of Int'l Conf. on Multimedia Modeling (MMM)*, pages 267–278. 2018.

[7] X. Fu, J. Yan, and C. Fan. Image Aesthetics Assessment Using Composite Features from off-the-Shelf Deep Models. In *Proc. of Int'l Conf. on Image Processing (ICIP)*, pages 3528–3532, 2018.

[8] L. Hou, C.-P. Yu, and D. Samaras. Squared Earth Mover's Distance-based Loss for Training Deep Neural Networks. *arXiv preprint arXiv:1611.05916*, 2016.

[9] B. Jin, M. V. O. Segovia, and S. Susstrunk. Image aesthetic predictors based on weighted CNNs. In *Proc. of Int'l Conf. on Image Processing (ICIP)*, pages 2291–2295, 2016.

[10] X. Jin, L. Wu, X. Li, S. Chen, S. Peng, J. Chi, S. Ge, C. Song, and G. Zhao. Predicting Aesthetic Score Distribution through Cumulative Jensen-Shannon Divergence. In *Proc. of AAAI Conf. on Artificial Intelligence (AAAI)*, 2018.

[11] Y. Kao, R. He, and K. Huang. Deep Aesthetic Quality Assessment With Semantic Information. *IEEE Trans. on Image Processing (TIP)*, 26(3):1482–1495, 2017.

[12] Y. Kao, C. Wang, and K. Huang. Visual aesthetic quality assessment with a regression model. In *Proc. of Int'l Conf. on Image Processing (ICIP)*, pages 1583–1587, 2015.

[13] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 419–426, 2006.

[14] K. Ko, J.-T. Lee, and C.-S. Kim. PAC-Net: Pairwise Aesthetic Comparison Network for Image Aesthetic Assessment. In *Proc. of Int'l Conf. on Image Processing (ICIP)*, pages 2491–2495, 2018.

[15] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo Aesthetics Ranking Network with Attributes and Content Adaptation. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 662–679, 2016.

[16] E. Levina and P. Bickel. The Earth Mover's distance is the Mallows distance: some insights from statistics. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, volume 2, pages 251–256, 2001.

[17] K. Y. Lo, K. H. Liu, and C. S. Chen. Assessment of photo aesthetics with efficiency. In *Proc. of Int'l Conf. on Pattern Recognition (ICPR)*, pages 2186–2189, 2012.

[18] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. RAPID: Rating Pictorial Aesthetics using Deep Learning. In *Proc. of ACM Int'l Conf. on Multimedia (ACMMM)*, volume 137, pages 457–466, 2014.

[19] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rating Image Aesthetics Using Deep Learning. *IEEE Trans. on Multimedia (TMM)*, 17(11):2021–2034, 2015.

[20] 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. of Int'l Conf. on Computer Vision (ICCV)*, pages 990–998, 2015.

[21] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 2206–2213, 2011.

[22] S. Ma, J. Liu, and C. Wen Chen. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 4535–4544, 2017.

[23] L. Mai, H. Jin, and F. Liu. Composition-Preserving Deep Photo Aesthetics Assessment. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 497–506, 2016.

[24] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 2408–2415. IEEE, 2012.

[25] A. Paszke, G. Chanan, Z. Lin, S. Gross, E. Yang, L. Antiga, and Z. Devito. Automatic differentiation in PyTorch. In *Proc. of Neural Information Processing Systems Workshops (NIPS Workshops)*, pages 1–4, 2017.

[26] Preferred Networks Inc. Optuna. from `https://optuna.org/`.

[27] H. Roy, T. Yamasaki, and T. Hashimoto. Predicting Image Aesthetics using Objects in the Scene. In *Proc. of Int'l Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia (MMArt&ACM')*, pages 14–19, 2018.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *Int'l Jour. of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[29] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[30] K. Schwarz, P. Wieschollek, and H. P. A. Lensch. Will People Like Your Image? Learning the Aesthetic Space. In *Proc. of Winter Conf. on Applications of Computer Vision (WACV)*, pages 2048–2057. IEEE, 2018.

[31] C. Shen, Z. Jin, Y. Zhao, Z. Fu, R. Jiang, Y. Chen, and X.-S. Hua. Deep Siamese Network with Multi-level Similarity Perception for Person Re-identification. In *Proc. of ACM Int'l Conf. on Multimedia (ACMMM)*, volume 17, pages 1942–1950, 2017.

[32] K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang, and B.-G. Hu. Attention-based Multi-Patch Aggregation for Image Aesthetic Assessment. In *Proc. of ACM Int'l Conf. on Multimedia (ACMMM)*, pages 879–886, 2018.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[34] H. Talebi and P. Milanfar. NIMA: Neural Image Assessment. *IEEE Trans. on Image Processing (TIP)*, 27(8):3998–4011, 2018.

[35] Z. Wang, D. Liu, S. Chang, F. Dolcos, D. Beck, and T. Huang. Image aesthetics assessment using Deep Chatterjee's machine. In *Proc. of Int'l Joint Conf. on Neural Networks (IJCNN)*, pages 941–948, 2017.

[36] Y. Zhou, X. Lu, J. Zhang, and J. Z. Wang. Joint Image and Text Representation for Aesthetics Analysis. In *Proc. of ACM Int'l Conf. on Multimedia (ACMMM)*, pages 262–266, 2016.