Computational Visual Media
https://doi.org/10.1007/s41095-021-0207-y Vol. 7, No. 2, June 2021, 241–252

Research Article

Joint regression and learning from pairwise rankings for personalized image aesthetic assessment

Jin Zhou[1], Qing Zhang[2] (✉), Jian-Hao Fan[2], Wei Sun[1], and Wei-Shi Zheng[2,3,4]

Abstract Recent image aesthetic assessment methods have achieved remarkable progress due to the emergence of deep convolutional neural networks (CNNs). However, these methods focus primarily on predicting generally perceived preference of an image, making them usually have limited practicability, since each user may have completely different preferences for the same image. To address this problem, this paper presents a novel approach for predicting personalized image aesthetics that fit an individual user's personal taste. We achieve this in a coarse to fine manner, by joint regression and learning from pairwise rankings. Specifically, we first collect a small subset of personal images from a user and invite him/her to rank the preference of some randomly sampled image pairs. We then search for the K-nearest neighbors of the personal images within a large-scale dataset labeled with average human aesthetic scores, and use these images as well as the associated scores to train a generic aesthetic assessment model by CNN-based regression. Next, we fine-tune the generic model to accommodate the personal preference by training over the rankings with a pairwise hinge loss. Experiments demonstrate that our method can effectively learn personalized image aesthetic preferences, clearly outperforming state-of-the-art methods. Moreover, we show that the learned personalized image aesthetic benefits a wide variety of applications.

Keywords personalized image aesthetic assessment; deep convolutional neural networks; pairwise ranking; regression

1 Introduction

The explosive growth of digital images has spawned automatic image aesthetic assessment, which is an important research problem that benefits a wide variety of applications, including photo album management, automatic image enhancement, image retrieval, and media recommendation. Despite being studied for decades, this problem remains a challenge because of the inherent uncertainty and subjectivity. While recent learning-based methods have made remarkable progress by leveraging the advantage of CNNs in scene understanding and feature learning, they are mostly designed for learning a universal image aesthetic assessment model that represents the average preference. However, in most applications of image aesthetics, e.g., automatic image/video beautification and recommendation [1–7], the user's personal preference for an image is usually more desirable than average preference, since different users may have substantially different preferences for the same image, as demonstrated in Fig. 1.

Compared to generic or universal image aesthetic assessment, personalized image aesthetic assessment is a more challenging problem. Large-scale datasets (e.g., AVA dataset [8] and AADB dataset [9]) labeled with average human ratings or attributes already exist for generic aesthetic model training. In contrast, it is usually impractical to collect a large number of personal images labeled with the owner's visual preference, since not everyone maintains a large photo

1 School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China. E-mail: J. Zhou, zhouj289@mail2.sysu.edu.cn; W. Sun, sunwei@mail.sysu.edu.cn.

2 School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China. E-mail: Q. Zhang, zhangqing.whu.cs@gmail.com (✉); J.-H. Fan, fanjh8@mail2.sysu.edu.cn; W.-S. Zheng, wszheng@ieee.org.

3 Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education (Sun Yat-sen University), Guangzhou 510006, China.

4 Peng Cheng Laboratory, Shenzhen 518000, China

Manuscript received: 2021-01-04; accepted: 2021-01-23

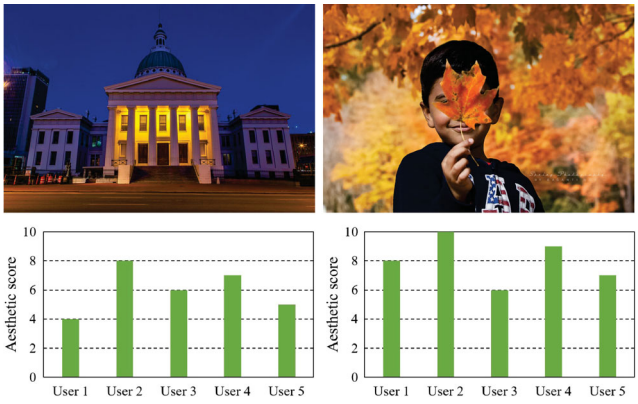


Fig. 1 Two example images rated by five different users using aesthetic scores from 1 (worst) to 10 (best). Above: input images. Below: rating distributions of the two images. Ratings are inconsistent across users over the two images: different users may have completely different visual preferences for the same image.

album, and rating image aesthetics could be tedious and unreliable for a single human agent.

Some research efforts have been made to tackle the personalized image aesthetic assessment problem. Ren et al. [10] proposed a residual-based model for accommodating individual aesthetic taste, while Park et al. [11] integrated personal preference into a generic aesthetic model by training a support vector machine (SVM) over pairwise ranking information. More recently, collaborative filtering [12, 13] has been employed to assess personal aesthetic preference [14, 15]. Despite the notable progress achieved by these methods, they still have limitations. Firstly, they usually collect absolute preference ratings of each personal image from the user, but we have found that such ratings are often unreliable since it is extremely difficult for a person to explicitly quantify his/her visual preference into discrete rating levels (e.g., the commonly adopted 1 to 10). Secondly, these methods may fail to effectively learn personalized aesthetic preferences from limited personal data.

In this paper, we present a novel personalized image aesthetic assessment method that is able to address these limitations of previous methods. Firstly, instead of directly collecting absolute aesthetic scores, we argue that it is more practical and reliable to collect relative preference rankings between images, since it is usually much easier and more reliable for a person to tell which one of two images he/she prefers, than to rate a single image with an absolute score. Thus, we ask the user to state his/her preference for a small number of image pairs randomly sampled from the collected personal images. Next, we enrich the pairwise ranking information by inferring new rankings from user-annotated rankings based on ranking transitivity, which largely remedies the lack of labeled data and allows us to more effectively learn personal preferences. Specifically, our approach comprises two stages. We first search for K-nearest neighbors of the collected personal images within a public aesthetic annotated benchmark dataset, and then train a generic aesthetic model from the searched images and the corresponding aesthetic scores using CNN-based regression. Finally, we adjust the generic model to fit the personal preference by learning from the pairwise rankings with a hinge loss.

In summary, the major contributions of this work are:

- a novel approach for learning personalized image aesthetics from very limited personal data, by joint regression and learning from rankings;
- extensive experiments to evaluate the proposed approach and compare it with various existing methods; results show that our method can more effectively learn personal aesthetic preferences;
- a demonstration that our learned personalized image aesthetics can be applied to customizing image retouching applications for a specific user, including exposure correction, color enhancement, and image dehazing.

2 Related work

2.1 Generic image aesthetic assessment

Most existing image aesthetic assessment methods aim to learn a generic aesthetic model based on the assumption that an implicit consensus exists about perceptually pleasant images. Early works treat image aesthetic prediction as a classification or regression problem of directly mapping hand-crafted visual features to aesthetic ratings provided by human raters [16–18]. With the emergence of large-scale aesthetic analysis datasets and deep neural networks, significant progress has been made towards automatic aesthetic assessment. Lu et al. [19] presented a multi-patch aggregation network for aesthetic classification, which was then improved in Ref. [20] to incorporate a visual attention mechanism. Mai et al. [21] introduced a scene-aware network with adaptive spatial pooling to learn image aesthetics. Kong et al. [9] achieved

photo aesthetics ranking by jointly learning image attribute and content information. Talebi and Milanfar [22] predicted the distribution of aesthetic scores using a convolutional neural network. Zeng et al. [23] presented a unified probabilistic formulation for image aesthetic assessment, while Zhang et al. [24] achieved unified aesthetic prediction through a gated peripheral-foveal convolutional neural network. More recently, Pan et al. [25] developed an image aesthetic assessment model assisted by attributes through adversarial learning. Wang et al. [26] devised a non-reference image quality assessment method for synthetic images based on convolutional neural networks and local image saliency. Sheng et al. [27] proposed the use of self-supervised feature learning for aesthetic prediction. See Ref. [28] for a survey of generic image aesthetic assessment.

2.2 Personalized image aesthetic assessment

Recently, there has been some research efforts towards personalized image aesthetic assessment. Ren et al. [10] achieved the goal by exploring the correlation between individual user's preferences and generic aesthetic perception, while Park et al. [11] adopted ranking information between images to train an SVM to predict personal preferences. Another main line of research is to use collaborative filtering, a fundamental algorithm used by recommendation systems to produce personal recommendations, for personalized aesthetic prediction. Following this line, Wang et al. [14] devised a deep aesthetic assessment model that integrates collaborative and attentive learning, while Korhonen [15] predicted personally perceived image quality by combining classical image feature analysis and collaboration filtering. In contrast to methods built upon collaborative filtering, Li et al. [29] designed a personality driven multi-task deep model for this purpose. Lee and Kim [30] used eigenvalue decomposition of a pairwise comparison matrix that involves multiple reference images and an input image. More recently, Zhu et al. [31] addressed the problem via meta-learning with bi-level gradient optimization, while Cui et al. [32] proposed to infer users' personal preferences based on their favoring behavior on social media platforms.

2.3 Learning to rank

Learning to rank has recently emerged as an attractive technique to train models for various vision and multimedia tasks. Yan et al. [33] trained a ranking model based on multiple additive regression trees for automatic image color enhancement. Paisitkriangkrai et al. [34] exploited learning to rank in person re-identification with metric ensembles. Liu et al. [35] used learning from rankings as a data augmentation technique for non-reference image quality assessment. Liu et al. [36] employed unlabeled data for crowd counting by learning to rank. In addition to the abovementioned application scenarios, learning to rank has also been applied to multi-label image classification [37, 38].

3 Our approach

This section describes our personalized image aesthetic assessment approach. We first introduce how we collect pairwise preference rankings using personal images collected from a specific user. Next, we associate the collected personal images with AVA—currently the largest public aesthetic analysis dataset—and train a generic aesthetic model. Finally, we illustrate how we adjust the generic model with the pairwise rankings to accommodate personal taste, and consider implementation details. An overview of our approach is shown in Fig. 2.

3.1 Personal data collection

To collect the personal data, we first invited a user to share with us a small set of personal images. The user was asked to carefully selected the images to have diverse contents, styles, lighting conditions, and colors. Then, the user was asked to provide a pairwise preference for some randomly sampled image pairs based on the user interface shown in Fig. 3. Unlike previous methods which require the users to perform many pairwise rankings, we found that it is feasible to infer many useful rankings from user-annotated rankings with the Floyd–Warshall algorithm based on ranking transitivity. For instance, for three personal images I_1, I_2, and I_3, if the user-annotated rankings are $I_1 > I_2$ and $I_2 > I_3$, then we can generate $I_1 > I_3$ by transitivity. Note that each newly generated pairwise ranking is associated with only two user-annotated rankings to avoid loops and to maintain reliability of the generated rankings. In other words, we do not generate $I_1 > I_4$, even we have $I_1 > I_2$, $I_2 > I_3$, and $I_3 > I_4$.

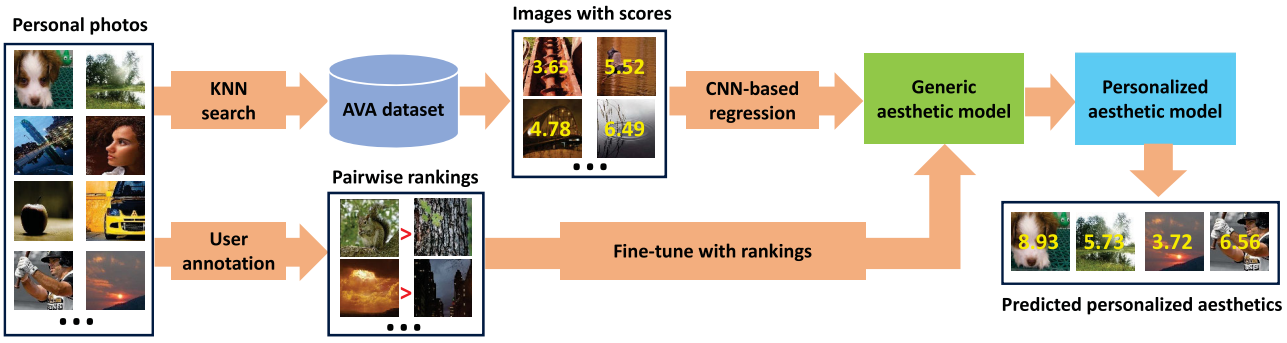


Fig. 2 Overview. Given a collection of a user's personal images, we first find each image's K-nearest neighbors (KNN) within the public AVA dataset. The discovered images and the corresponding aesthetic scores are then fed into a CNN-based regression network to train a generic aesthetic model. Next, we asked the user to state a preference in a few image pairs sampled from the personal image collection, and infered new rankings based on ranking transitivity. Finally, all pairwise ranking information is used to fine-tune the generic aesthetic model, turning it into a personalized aesthetic model that fits personal taste.

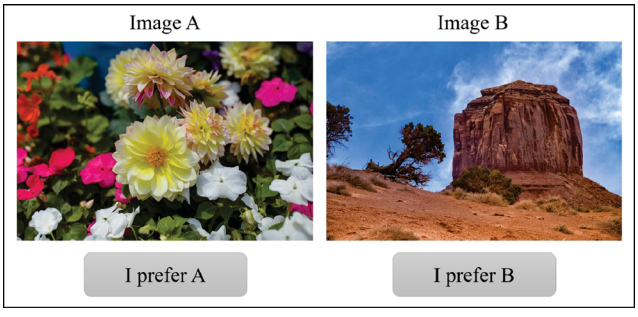


Fig. 3 User interface for collecting pairwise rankings. For each image pair, the user was asked to select the image that he/she prefers. The preferred image is ranked higher.

3.2 Generic image aesthetic regression

With the collected personal images, we regress a generic aesthetic model that numerically describes the universal visual preference for images of similar categories. To this end, we first perform a KNN search for each personal image within the AVA dataset. The discovered images and the corresponding aesthetic scores are then employed to train a generic aesthetic model via CNN-based regression. Below we describe the above two components, KNN image searching and CNN-based aesthetic regression, in detail.

3.2.1 KNN image searching

To obtain the KNN, we first obtain normalized feature vectors for each personal image and the images from the AVA dataset, based on VGG16 [39] pre-trained on ImageNet [40]. Next, we search for the KNN of each personal image from the AVA dataset by measuring the cosine distance between corresponding feature vectors. In our experiment, we empirically set $K = 50$, since it not only ensures that we collect sufficient training data for CNN-based aesthetic regression, but also allows more efficient network training.

3.2.2 CNN-based aesthetic regression

Figure 4 shows the overall network architecture of our CNN-based aesthetic regression network. Specifically, VGG16 is utilized to extract feature maps, which consist of 16 layers, 13 convolutional layers with small convolution filters of size 3×3, and 3 fully connected layers. To allow input of images of arbitrary size and back-propagation from aesthetic scores to original pixels, we remove the last three fully connected layers in the original VGG16 and add a max-pooling layer. For a given image, we first extract three feature maps Z_1, Z_2, and Z_3 from VGG16, which are then fed into two convolutional attention modules to get the attention maps A_1 and A_2. Next, the predicted attention maps operate on the features in Z_1 and Z_2 via point-wise multiplication. This design is inspired by the physiological observation that local contexts typically play a more important role in visual preference evaluation at first glance. Finally, the attentive features are concatenated and fed into a fully connected layer with 10 neurons (annotated ratings are 1 to 10 for images in the AVA dataset) to predict the actual aesthetic score.

Now we describe the training loss for the CNN-based aesthetic regression network. Each image in the AVA dataset is assigned a set of user ratings ranging from 1 to 10 in terms of empirical probability mass function $\boldsymbol{p} = [p_1, \ldots, p_{10}]$, $\sum_{i=1}^{10} p_i = 1$, where p_i, $i \in [1, 10]$ denotes the probability the image is labeled with aesthetic score i. Our goal is to predict the probability distribution of aesthetic scores for a

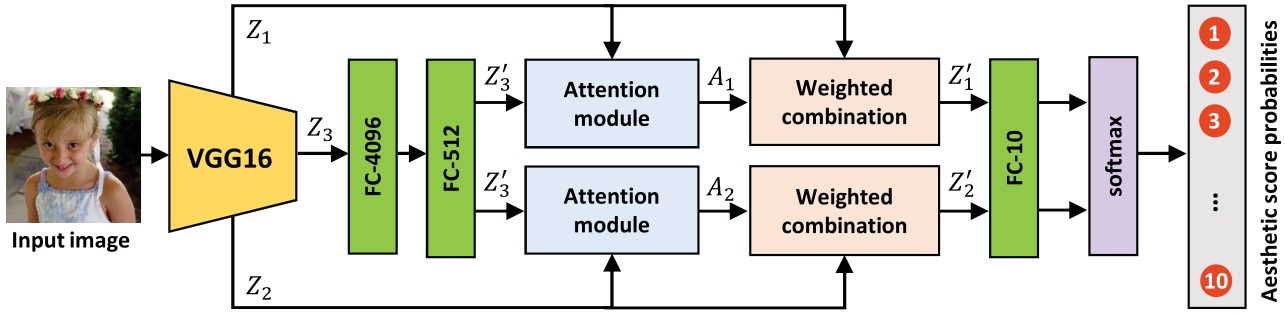


Fig. 4 Network architecture of our CNN-based aesthetic regression network. Given an input image, we first send it into VGG16 to get three feature maps, i.e., Z_1: 10th convolutional (conv) layer, Z_2: 13th conv layer, and Z_3: 13th conv layer with max-pooling. The feature maps (Z_1, Z_2, and Z_3') are then fed into two attention modules to get the attention maps A_1 and A_2, which are used to generate Z_1' and Z_2' by weighted combination with Z_1 and Z_2. Finally, we concatenate Z_1' and Z_2' to form the final feature representation, and employ a fully connected layer with 10 output neurons followed by the softmax function to predict the aesthetic score probabilities for the input image.

given image. To regress a generic aesthetic model based on the discovered images and their associated rating annotations, we employ the Earth Mover's Distance (EMD) to formulate a loss for network training. It performs well due to its ability to penalize misclassifications according to class distance. Formally, the loss is defined as

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{10}\sum_{j=1}^{10}|\mathcal{C}_{\boldsymbol{p}}(j) - \mathcal{C}_{\hat{\boldsymbol{p}}}(j)|^{\ell}\right)^{1/\ell} \quad (1)$$

where N denotes the total number of discovered images. $\mathcal{C}_{\boldsymbol{p}}(j) = \sum_{i=1}^{j} p_i$ denotes the cumulative distribution function. $\hat{\boldsymbol{p}}$ denotes the probability mass function that we aim to estimate. ℓ is set to 2 to allow efficient optimization. Intuitively, the EMD-based loss measures the cost of moving the ground-truth distribution $\boldsymbol{p}$ to the estimated distribution $\hat{\boldsymbol{p}}$. The mean score obtained from the estimated distribution $\hat{\boldsymbol{p}}$ is used as the output aesthetic score, i.e., $\sum_{i=1}^{10} i\hat{p}_i$.

3.3 Personalized fine-tuning with pairwise rankings

Having obtained the generic aesthetic model, the next step is to incorporate personal visual preferences by fine-tuning the generic model according to the collected pairwise rankings. To do so, we retrain the CNN-based regression network with rankings by using a pairwise ranking hinge loss defined as

$$\mathcal{L}_h(x_1, x_2; \theta) = \max(0, f(x_2;\theta) - f(x_1;\theta) + \epsilon) \quad (2)$$

where x_1 and x_2 are a pair of images. θ denotes the network parameters. $f(x_1;\theta)$ and $f(x_2;\theta)$ represent the predicted aesthetic scores of images x_1 and x_2. ϵ is the margin, which is set to 0.1 in our experiments. Following Ref. [35], we assume without loss of generality that x_1 has higher score than x_2, so the gradient of the loss in Eq. (2) can be written as

$$\nabla_\theta \mathcal{L}_h = \begin{cases} 0, & \text{if } f(x_2;\theta) - f(x_1;\theta) + \epsilon \leqslant 0 \\ \nabla_\theta f(x_2;\theta) - \nabla_\theta f(x_1;\theta), & \text{otherwise} \end{cases} \quad (3)$$

The above equation implies that when the predicted scores of the network are in accordance with the pairwise ranking, the gradient is zero. While the pairwise ranking is not met, the gradient of the image with higher score (x_1) is decreased and the gradient of the other (x_2) will be increased.

3.4 Implementation details

Our network was implemented in TensorFlow, and optimized by the Adam optimizer. For the CNN-based aesthetic regression, we trained for 10 epochs with a batch size of 32 and a fixed learning rate of 3×10^{-7}. For the ranking-based fine-tuning stage, we trained for another 20 epochs with an initial learning rate of 5×10^{-6}. An exponential decay of 0.5 was applied to the learning rate after every 500 iterations.

4 Experiments

In this section, we describe experiments used to validate the effectiveness of the proposed approach. We first introduce the test datasets and evaluation metrics, and then compare our method to state-of-the-art methods. Next, we provide an in-depth analysis of our approach. Finally, we showcase several applications enabled by our approach.

4.1 Datasets

The benchmark AVA dataset [8] and the REAL-CUR dataset [10] were employed to evaluate our method.

The AVA dataset consists of 255,000 images, each of which is aesthetically rated by an average of 210 users with scores ranging from 1 to 10. The REAL-CUR dataset contains 14 personal photo albums (each one including about 200 images), and each personal image is annotated with aesthetic score ranging from 1 to 5. To unify the range of scores to [1, 10], the annotated aesthetic scores of the REAL-CUR dataset were doubled. The REAL-CUR dataset has the following two usages. Firstly, it provides the desired personal images for network training. Secondly, it can be used to verify the effectiveness of the learned personalized aesthetics. Specifically, we divided each album into two subsets, i.e., a set consisting of X images for network training and the other set containing the remaining personal images for testing. Then, we found the KNN for each image in the training subset from the AVA dataset to construct the regression training dataset. Next, we randomly selected 100 image pairs from each training subset, and got their pairwise rankings according to the annotated aesthetic scores (equal scores are discarded). Finally, the obtained pairwise rankings were enriched with the Floyd–Warshall algorithm, and we trained a personalized image aesthetic assessment model based on the regression training dataset and the collected pairwise rankings.

4.2 Evaluation metrics

Akin to prior methods [9, 31], ranking correlations are used to measure consistency between predictions and ground truth user scores. Specifically, we employed the Spearman rank-order correlation coefficient (ρ) [41] to quantitatively evaluate the performance of personalized image aesthetic assessment. It is defined as

$$\rho = 1 - \frac{6\sum_{j=1}^{M}(r_j - r'_j)^2}{M(M^2-1)} \tag{4}$$

where r_j denotes the rank of the jth test image when sorting the ground truth aesthetic scores in descending order, while r'_j denotes the rank given by the predicted aesthetic scores. M is the number of test images. The value of ρ ranges from -1 to 1, and a higher absolute value indicates stronger correlation and better overall performance.

4.3 Comparison with existing methods

4.3.1 Method

We compare our method with eight existing methods, including: NIMA [42], MPADA [20], MLSP [43], FPMF [44], PAM [10], as well as three other ranking-based methods: R-SVM [45], R-SVR [11], and RankIQA [35]. Note, the original RankIQA collects rankings by randomly distorting the input images for image quality assessment. To make it fit our task, we replaced their ranking data with our collected rankings. For fair comparison, we retrained the compared methods based on the images discovered from the AVA dataset and the collected pairwise rankings, using the publicly-available implementation provided by the authors with recommended parameter settings. We implemented R-SVR ourselves since there is no publicly available implementation. We did not compare with Ref. [14], since it relies on both personal ratings and image reviews for model training. Our comparison is twofold.

4.3.2 Quantitative comparison

Table 1 reports a quantitative comparison of our method with the other methods, using 10 ($X = 10$) and 100 ($X = 100$) training images, respectively. The mean ranking correlation of all 14 personal photo albums in REAL-CUR is shown in Table 1. As can be seen, directly learning the personal visual preference from very limited training data via naive regression (NIMA) or collaborative filtering (FPMF) results in poor generalizability to unseen test images. PAM produces very competitive results by simultaneously

Table 1 Quantitative comparison between our method and state-of-the-art methods on the REAL-CUR dataset in terms of rank correlation ρ. Exp. loss and Log. loss stand for exponential loss and logistic loss, alternatives to the hinge loss in Eq. (2). Pre-training indicates the CNN-based generic aesthetic regression on AVA. w/o means without

Method	10 images	100 images
NIMA [22]	0.41 ± 0.13	0.43 ± 0.01
MPADA [20]	0.46 ± 0.11	0.48 ± 0.02
MLSP [43]	0.43 ± 0.12	0.45 ± 0.01
FPMF [44]	0.37 ± 0.14	0.40 ± 0.01
PAM [10]	0.58 ±0.10	0.65 ± 0.02
R-SVM [45]	0.33 ± 0.12	0.39 ± 0.03
R-SVR [11]	0.47 ± 0.11	0.55 ± 0.02
RankIQA [35]	0.56 ± 0.12	0.63 ± 0.01
Ours with Exp. loss	0.55 ± 0.13	0.61 ± 0.02
Ours with Log. loss	0.54 ± 0.12	0.58 ± 0.01
Ours w/o attention	0.53 ± 0.14	0.57 ± 0.03
Ours w/o rankings	0.44 ± 0.10	0.49 ± 0.02
Ours w/o pre-training	0.25 ± 0.11	0.31 ± 0.01
Our full method	**0.61 ± 0.11**	**0.67 ± 0.02**

considering content and aesthetic attributes. Pairwise rankings are adopted in the three compared ranking-based methods (R-SVM, R-SVR, and RankIQA), yet our method outperforms them, showing that our method not only effectively learns personalized aesthetics from very limited data but also generalizes well to unseen personal images. Figure 5 compares personalized aesthetic scores predicted by our method and the comparative methods on some example test images from the REAL-CUR dataset. As can be seen, our personalized aesthetic assessment model more accurately predicts the user's ratings.

4.3.3 User study

As personalized aesthetic assessment is highly subjective, we further conducted a user study with 4 users (2 males and 2 females) to evaluate our method. To this end, we first collected four personal image datasets from the users, covering a broad range of scenes, subjects, and lighting conditions. The four personal datasets are referred to as PD1, ..., PD4, and each contains 200 images. We then randomly selected 150 personal images from each dataset and collected 220 pairwise rankings among these images from the corresponding user, while the remaining 50 personal images were reserved for testing. Next, we trained personalized aesthetic assessment models using our approach and the other three ranking-based methods (R-SVM, R-SVR, and RankIQA), and used the trained models to predict aesthetic scores for all testing images. To assess performance, we randomly selected image pairs from the test images and showed the corresponding user the personalized aesthetic scores predicted by different methods, and asked the user to judge whether the rankings indicated by the predicted scores were consistent with his/her personalized visual preference. Table 2 summarizes the percentage of pairwise rankings predicted by different methods that are consistent with the specific personalized user preference. We can see that our predicted aesthetic scores better match the user's preference. Figure 6 shows some example results for image pairs employed in the user study.

4.4 Ablation study

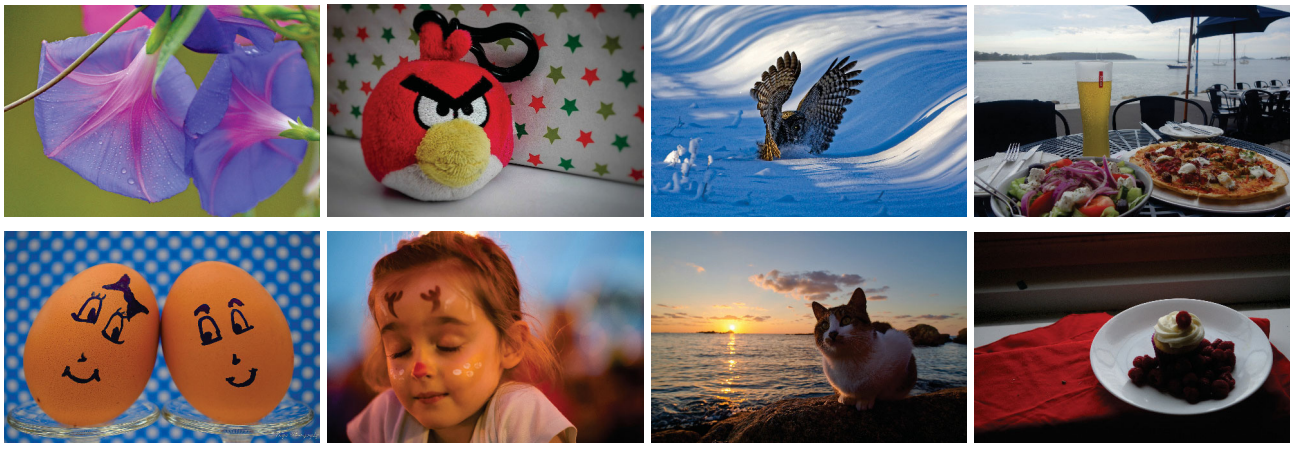
We also quantitatively evaluated the effectiveness of the CNN-based regression, the learning from ranking

Table 2 Percentage of predicted pairwise rankings consistent with the user preference, for our method and three state-of-the-art ranking-based methods on four users' personal datasets

Method	PD1	PD2	PD3	PD4
R-SVM [42]	67.1%	69.5%	61.3%	65.7%
R-SVR [11]	69.5%	72.6%	70.4%	73.8%
RankIQA [35]	85.1%	86.4%	84.4%	85.3%
Ours	**91.3%**	**87.3%**	**86.9%**	**88.5%**

Method	S	Method	S	Method	S	Method	S	Method	S	Method	S
NIMA	4.81	NIMA	4.43	NIMA	5.11	NIMA	8.34	NIMA	8.23	NIMA	7.45
MPADA	3.84	MPADA	3.64	MPADA	5.24	MPADA	7.34	MPADA	8.84	MPADA	7.91
MLSP	4.12	MLSP	3.92	MLSP	4.92	MLSP	7.91	MLSP	8.91	MLSP	8.15
FPMF	5.42	FPMF	5.17	FPMF	4.61	FPMF	4.18	FPMF	6.37	FPMF	6.37
PAM	3.20	PAM	1.38	PAM	7.26	PAM	7.37	PAM	9.41	PAM	9.76
R-SVM	3.63	R-SVM	3.71	R-SVM	5.18	R-SVM	3.64	R-SVM	7.62	R-SVM	6.51
R-SVR	4.75	R-SVR	5.23	R-SVR	8.42	R-SVR	5.21	R-SVR	9.33	R-SVR	7.45
RankIQA	1.12	RankIQA	3.34	RankIQA	6.54	RankIQA	8.73	RankIQA	8.75	RankIQA	8.96
Ours	2.36	Ours	2.18	Ours	5.86	Ours	6.25	Ours	9.81	Ours	9.62
Owner	2.0	Owner	2.0	Owner	6.0	Owner	6.0	Owner	10	Owner	10

Fig. 5 Comparison of personalized aesthetic scores S predicted by our method and state-of-the-art methods, for some test images from the REAL-CUR dataset. Above: test images. Below: personalized aesthetic scores predicted by different methods and ground truth scores given by the image owner. The predicted aesthetic score closest to the user-labeled score is highlighted by a gray beckgroun.



User	Top > Bottom	User	Top < Bottom	User	Top < Bottom	User	Top > Bottom
R-SVM	✗	R-SVM	✗	R-SVM	✗	R-SVM	✗
R-SVR	✓	R-SVR	✗	R-SVR	✗	R-SVR	✓
RankIQA	✗	RankIQA	✗	RankIQA	✓	RankIQA	✓
Ours	✓	Ours	✓	Ours	✓	Ours	✓

Fig. 6 Example results on four image pairs (columns 1–4) employed in the user study. The tables show the user-labeled preference ranking (top row) for the corresponding image pair and the ranking predicted by different methods (rows 2–5). ✓ indicates a predicted ranking consistent with the user's taste, and ✗ indicates an inconsistent result.

design, and the attention mechanism in the regression network. Comparing the 2nd row and 13th row in Table 1, we observe a clear advantage of our regression network over a baseline regression network (NIMA). Moreover, in addition to our utilized pairwise ranking hinge loss, we also tried two other commonly used alternatives, exponential loss and logistic loss [46]. As shown, using the same regression network, the hinge loss achieved better results than the other two losses, convincingly demonstrating its effectiveness. As can be observed by comparing the 12th and 14th rows with the 15th row, omitting the attention module from the regression network and the pre-training on AVA leads to an obvious decrease in overall performance, demonstrating that they are beneficial to learning personalized aesthetics.

4.5 Limitations

Our method may fail to accurately predict personal preferences when the collected personal images are severely imbalanced in subjects and scenes. For instance, when most personal images belong to a single category (e.g., indoor images), our method may fail to predict the individual's preferences for other kinds of images (e.g., portraits).

4.6 Applications

Our approach can be applied to personalized image retouching to better meet users' personalized tastes. To do so, we designed an aesthetic quality loss $\mathcal{L}_{\text{aesthetic}}(x) = 10 - f(x)$, where x and $f(x)$ denote the retouched image and the predicted personalized aesthetic score (f denotes our trained personalized aesthetic assessment model). Intuitively, this loss enforces the score of the retouched image to be as close to the maximum (10) as possible. By incorporating the loss for training a specific learning-based image retouching framework, we can achieve personalized image retouching. Figures 7–9 show the use of our learned personalized aesthetic for a user who favors bright scenes, vivid colors, and clear details in image retouching tasks of exposure correction, color enhancement, and image dehazing. As shown, incorporating personalized aesthetics produces results which better satisfy the user's preferences.

(a) Input (2.32) (b) DeepUPE (5.36) (c) DeepUPE+ (7.57)

Fig. 7 Personalized exposure correction. (a) Input. (b, c) Results from DeepUPE [47] and personalized DeepUPE (DeepUPE+). Aesthetic scores are shown in parentheses.

(a) Input (6.12) (b) DAR (7.76) (c) DAR+ (8.83)

Fig. 8 Personalized color enhancement. (a) Input. (b, c) Results from DAR [48] and personalized DAR (DAR+).

(a) Input (2.19) (b) DMMF (6.21) (c) DMMF+ (6.85)

Fig. 9 Personalized image dehazing. (a) Input. (b, c) Results from DMMF [49] and personalized DMMF (DMMF+).

5 Conclusions

We have presented a novel approach for personalized image aesthetic assessment. Unlike previous methods that are mostly based on user-annotated absolute aesthetic ratings, we distill an individual user's visual preference by joint regression and learning from pairwise rankings, which not only allows more accurate aesthetic learning, but also remedies the lack of labeled data. We first collect a small set of personal images and find their K nearest neighbors from the benchmark AVA dataset, and then train a generic aesthetic model with the discovered aesthetic labeled images. Next, we adjust the generic model to accommodate personal taste by incorporating user annotated ranking information. Experiments demonstrate the effectiveness of our method.

Acknowledgements

The authors thank the reviewers for their valuable comments. This work was supported partially by the National Key Research and Development Program of China (2018YFB1004903), National Natural Science Foundation of China (61802453, U1911401, U1811461), Fundamental Research Funds for the Central Universities (19lgpy216), and Research Projects of Zhejiang Lab (2019KD0AB03).

References

[1] Zhang, F.-L.; Wang, M.; Hu, S.-M. Aesthetic image enhancement by dependence-aware object recomposition. *IEEE Transactions on Multimedia* Vol. 15, No. 7, 1480–1490, 2013.

[2] Zhang, Q.; Nie, Y. W.; Zhang, L.; Xiao, C. X. Underexposed video enhancement via perception-driven progressive fusion. *IEEE Transactions on Visualization and Computer Graphics* Vol. 22, No. 6, 1773–1785, 2016.

[3] Zhang, Q.; Yuan, G. Z.; Xiao, C. X.; Zhu, L.; Zheng, W. S. High-quality exposure correction of underexposed photos. In: Proceedings of the 26th ACM international conference on Multimedia, 582–590, 2018.

[4] Zhang, F. L.; Wu, X.; Li, R. L.; Wang, J.; Zheng, Z. H.; Hu, S. M. Detecting and removing visual distractors for video aesthetic enhancement. *IEEE Transactions on Multimedia* Vol. 20, No. 8, 1987–1999, 2018.

[5] Zhang, Q.; Nie, Y.; Zheng, W.-S. Dual illumination estimation for robust exposure correction. *Computer Graphics Forum* Vol. 38, 243–252, 2019.

[6] Zhang, Q.; Yin, G. L.; Nie, Y. W.; Zheng, W. S. Deep camouflage images. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 7, 12845–12852, 2020.

[7] Zhang, Q.; Nie, Y.; Zhu, L.; Xiao, C.; Zheng, W.-S. Enhancing underexposed photos using perceptually bidirectional similarity. *IEEE Transactions on Multimedia* Vol. 23, 189–202, 2021.

[8] Murray, N.; Marchesotti, L.; Perronnin, F. AVA: A large-scale database for aesthetic visual analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2408–2415, 2012.

[9] Kong, S.; Shen, X. H.; Lin, Z.; Mech, R.; Fowlkes, C. Photo aesthetics ranking network with attributes and content adaptation. In: Proceedings of the European Conference on Computer Vision, 662–679, 2016.

[10] Ren, J.; Shen, X.; Lin, Z.; Mech, R.; Foran, D. J. Personalized image aesthetics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 638–647, 2017.

[11] Park, K.; Hong, S.; Baek, M.; Han, B. Personalized image aesthetic quality assessment by joint regression and ranking. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 1206–1214, 2017.

[12] Sarwar, B.; Karypis, G.; Konstan, J.; Reidl, J. Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international Conference on World Wide Web, 285–295, 2001.

[13] Breese, J. S.; Heckerman, D.; Kadie, C. Empirical analysis of predictive algorithms for collaborative filtering. *arXiv preprint* arXiv:1301.7363, 2013.

[14] Wang, G.; Yan, J.; Qin, Z. Collaborative and attentive learning for personalized image aesthetic assessment. In: Proceedings of the International Joint Conference on Artificial Intelligence, 957–963, 2018.

[15] Korhonen, J. Assessing personally perceived image quality via image features and collaborative filtering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 8169–8177, 2019.

[16] Luo, W.; Wang, X.; Tang, X. Content-based photo quality assessment. In: Proceedings of the IEEE International Conference on Computer Vision, 2206–2213, 2011.

[17] Dhar, S.; Ordonez, V.; Berg, T. L. High level describable attributes for predicting aesthetics and interestingness. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1657–1664, 2011.

[18] Marchesotti, L.; Perronnin, F.; Larlus, D.; Csurka, G. Assessing the aesthetic quality of photographs using generic image descriptors. In: Proceedings of the IEEE International Conference on Computer Vision, 1784–1791, 2011.

[19] Lu, X.; Lin, Z.; Shen, X.; Mech, R.; Wang. J. Z. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In: Proceedings of the IEEE International Conference on Computer Vision, 990–998, 2015.

[20] Sheng, K. K.; Dong, W. M.; Ma, C. Y.; Mei, X.; Huang, F. Y.; Hu, B. G. Attention-based multi-patch aggregation for image aesthetic assessment. In: Proceedings of the ACM International Conference on Multimedia, 879–886, 2018.

[21] Mai, L.; Jin, H.; Liu, F. Composition-preserving deep photo aesthetics assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 497–506, 2016.

[22] Talebi, H.; Milanfar, P. NIMA: Neural image assessment. *IEEE Transactions on Image Processing* Vol. 27, No. 8, 3998–4011, 2018.

[23] Zeng, H.; Cao, Z.; Zhang, L.; Bovik, A. C. A unified probabilistic formulation of image aesthetic assessment. *IEEE Transactions on Image Processing* Vol. 29, 1548–1561, 2019.

[24] Zhang, X. D.; Gao, X. B.; Lu, W.; He, L. H. A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction. *IEEE Transactions on Multimedia* Vol. 21, No. 11, 2815–2826, 2019.

[25] Pan, B. W.; Wang, S. F.; Jiang, Q. S. Image aesthetic assessment assisted by attributes through adversarial learning. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 33, 679–686, 2019.

[26] Wang, X. C.; Liang, X. H.; Yang, B. L.; Li, F. W. B. No-reference synthetic image quality assessment with convolutional neural network and local image saliency. *Computational Visual Media* Vol. 5, No. 2, 193–208, 2019.

[27] Sheng, K. K.; Dong, W. M.; Chai, M. L.; Wang, G. H.; Zhou, P.; Huang, F. Y.; Hu, B.; Ji, R.; Ma, C. Revisiting image aesthetic assessment via self-supervised feature learning. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 4, 5709–5716, 2020.

[28] Deng, Y. B.; Loy, C. C.; Tang, X. O. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine* Vol. 34, No. 4, 80–106, 2017.

[29] Li, L. D.; Zhu, H. C.; Zhao, S. C.; Ding, G. G.; Jiang, H. Y.; Tan, A. Personality driven multi-task learning for image aesthetic assessment. In: Proceedings of the International Conference on Multimedia and Expo, 430–435, 2019.

[30] Lee, J. T.; Kim, C. S. Image aesthetic assessment based on pairwise comparison: A unified approach to score regression, binary classification, and personalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 1191–1200, 2019.

[31] Zhu, H.; Li, L.; Wu, J.; Zhao, S.; Ding, G.; Shi, G. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *IEEE Transactions on Cybernetics* https://doi.org/10.1109/TCYB.2020.2984670, 2020.

[32] Cui, C. R.; Yang, W. Y.; Shi, C.; Wang, M.; Nie, X. S.; Yin, Y. L. Personalized image quality assessment with social-sensed aesthetic preference. *Information Sciences* Vol. 512, 780–794, 2020.

[33] Yan, J.; Lin, S.; Kang, S. B.; Tang, X. A learning-to-rank approach for image color enhancement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2987–2994, 2014.

[34] Paisitkriangkrai, S.; Shen, C. H.; van den Hengel, A. Learning to rank in person re-identification with metric ensembles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1846–1855, 2015.

[35] Liu, X. L.; van de Weijer, J.; Bagdanov, A. D. RankIQA: Learning from rankings for no-reference image quality assessment. In: Proceedings of the IEEE International Conference on Computer Vision, 1040–1049, 2017.

[36] Liu, X. L.; van de Weijer, J.; Bagdanov, A. D. Leveraging unlabeled data for crowd counting by learning to rank. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7661–7669, 2018.

[37] Gong, Y. C.; Jia, Y. Q.; Leung, T.; Toshev, A.; Ioffe, S. Deep convolutional ranking for multilabel image annotation. *arXiv preprint* arXiv:1312.4894, 2013.

[38] Wang, Y. L.; Wang, S. H.; Tang, J. L.; Liu, H.; Li, B. X. PPP: Joint pointwise and pairwise image label prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6005–6013, 2016.

[39] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556, 2014.

[40] Deng, J.; Dong, W.; Socher, R.; Li, L. J.; Li, K.; Li, F. F. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 248–255, 2009.

[41] Myers, J. L.; Well, A.; Lorch, R. F. *Research Design and Statistical Analysis*. Routledge, 2010.

[42] Talebi, H.; Milanfar, P. Learned perceptual image enhancement. In: Proceedings of the IEEE International Conference on Computational Photography, 1–13, 2018.

[43] Hosu, V.; Goldlücke, B.; Saupe, D. Effective aesthetics prediction with multi-level spatially pooled features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 9375–9383, 2019.

[44] O'Donovan, P.; Agarwala, A.; Hertzmann, A. Collaborative filtering of color aesthetics. In: Proceedings of the Workshop on Computational Aesthetics, 33–40, 2014.

[45] Joachims, T. Optimizing search engines using clickthrough data. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 133–142, 2002.

[46] Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; Hullender, G. Learning to rank using gradient descent. In: Proceedings of the 22nd International Conference on Machine Learning, 89–96, 2005.

[47] Wang, R.; Zhang, Q.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; Jia, J. Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6849–6857, 2019.

[48] Park, J.; Lee, J. Y.; Yoo, D.; Kweon, I. S. Distort-and-recover: Color enhancement using deep reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5928–5936, 2018.

[49] Deng, Z.; Zhu, L.; Hu, X.; Fu, C.-W.; Xu, X.; Zhang, Q.; Qin, J.; Heng, P.-A. Deep multi-model fusion for single-image dehazing. In: Proceedings of the IEEE International Conference on Computer Vision, 2453–2462, 2019.

Jin Zhou is a master student in the School of Electronics and Information Technology, Sun Yat-sen University. His research interests include computer vision and deep learning.

Qing Zhang is a research associate professor in the School of Computer Science and Engineering, Sun Yat-sen University. His research interests include computational photography and computer vision.

Jian-Hao Fan is an undergraduate student in the School of Computer Science and Engineering, Sun Yat-sen University. His research interests are computer vision and deep learning.

Wei Sun received his Ph.D. degree in computer science from Sun Yat-sen University in 2004, where he is currently a professor in the School of Electronics and Information Technology. His research interests include multimedia forensics and signal processing.

Wei-Shi Zheng received his Ph.D. degree in applied mathematics from Sun Yat-sen University in 2008. He is now a full professor in the School of Computer Science and Engineering, Sun Yat-sen University. His research interests include person/object association and activity understanding in visual surveillance, and the related large-scale machine learning algorithm. He has more than 90 publications in leading journals (TPAMI, IJCV, TNN/TNNLS, TIP, PR) and conferences (ICCV, CVPR,

IJCAI, AAAI). He is an associate editor of the *Pattern Recognition Journal*. He has joined Microsoft Research Asia Young Faculty Visiting Programme and is a recipient of the Excellent Young Scientists Fund of the National Natural Science Foundation of China, and the Royal Society Newton Advanced Fellowship, UK.