Towards Artistic Image Aesthetics Assessment: a Large-scale Dataset and a New Method

Ran Yi[1]*, Haoyuan Tian[1], Zhihao Gu[1], Yu-Kun Lai[2], Paul L. Rosin[2]
[1]Shanghai Jiao Tong University, [2]Cardiff University
{ranyi,thy0210,ellery-holmes}@sjtu.edu.cn, {LaiY4,RosinPL}@cardiff.ac.uk

Abstract

Image aesthetics assessment (IAA) is a challenging task due to its highly subjective nature. Most of the current studies rely on large-scale datasets (e.g., AVA and AADB) to learn a general model for all kinds of photography images. However, little light has been shed on measuring the aesthetic quality of artistic images, and the existing datasets only contain relatively few artworks. Such a defect is a great obstacle to the aesthetic assessment of artistic images. To fill the gap in the field of artistic image aesthetics assessment (AIAA), we first introduce a large-scale AIAA dataset: Boldbrush Artistic Image Dataset (BAID), which consists of 60,337 artistic images covering various art forms, with more than 360,000 votes from online users. We then propose a new method, SAAN (Style-specific Art Assessment Network), which can effectively extract and utilize style-specific and generic aesthetic information to evaluate artistic images. Experiments demonstrate that our proposed approach outperforms existing IAA methods on the proposed BAID dataset according to quantitative comparisons. We believe the proposed dataset and method can serve as a foundation for future AIAA works and inspire more research in this field. Dataset and code are available at: https://github.com/Dreemurr-T/BAID.git

1. Introduction

With the ever-growing scale of online visual data, image aesthetic assessment (IAA) shows great potential in a variety of applications such as photo recommendation, image ranking and image search [6]. In recent years, image style transfer [9, 14, 19, 20, 26] and AI painting [15, 39] have become high-profile research areas. Users can easily generate artworks of numerous styles from websites and online applications, which has led to the explosion of artistic images online and the drastic increase in demand for automatically evaluating artwork aesthetics. We refer to this problem as **artistic image aesthetic assessment (AIAA)**.

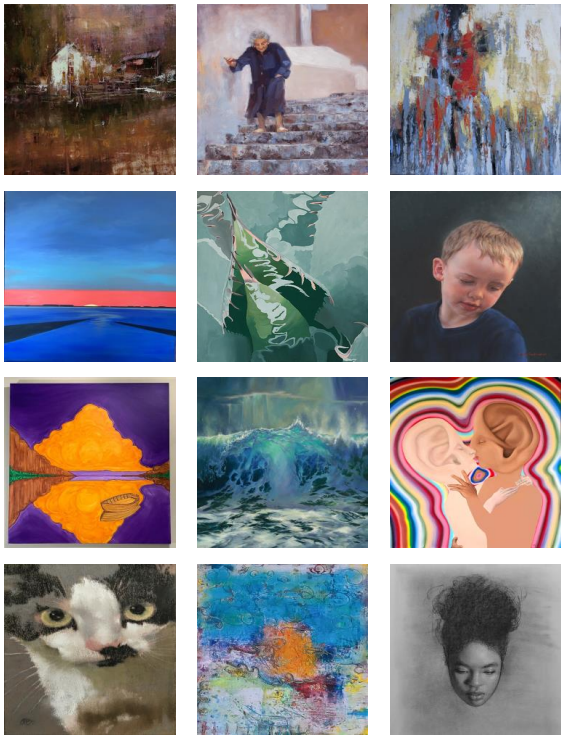


Figure 1. Samples from the proposed BAID dataset. BAID covers a wide range of artistic styles and painting themes.

The artistic image aesthetic assessment task is similar to IAA for being extremely challenging due to its highly subjective nature, as different individuals may have distinct visual and art preferences. Existing datasets related to this task can be summarized into three categories, but none of them meets the requirements of the AIAA task: (1) **IAA datasets**: modern IAA methods [13, 21, 23, 30, 32, 34] are data-driven, usually trained and evaluated on large-scale IAA datasets, *e.g.*, AVA [25], AADB [17] and CUHK-PQ [22]. However, these datasets only contain real-world

*Corresponding author.

photos and do not include artistic images like oil paintings or pencil sketches. This deficiency of artistic images is prevalent in existing IAA datasets [4, 16, 17, 22, 28], which means that given an artwork, existing IAA methods evaluate it based on perceptions learned from photography, and the evaluation is likely to be inaccurate since the perceptual rules of photography and art are not the same. (2) **Artistic datasets without aesthetic labels**: existing large-scale artistic image datasets [1, 29, 36] are mainly used to train style transfer, artistic style classification or text to image models, but they lack score annotations indicating image aesthetic level. (3) **Small-scale AIAA datasets**: efforts into building public AIAA datasets are scarce and the existing datasets [3, 8] contain relatively few number of images (less than 2,000). Based on the above observations, we conclude that **the lack of a large-scale AIAA dataset** is the biggest obstacle towards developing AIAA approaches.

To solve the problem, we first introduce a large-scale dataset specifically constructed for the AIAA task: the Boldbrush Artistic Image Dataset (BAID), which consists of 60,337 artworks annotated with more than 360,000 votes. The proposed BAID is, to our knowledge, the largest AIAA dataset, which far exceeds existing IAA and AIAA datasets in the quantity and quality of artworks.

Furthermore, we propose a baseline model, called the Style-specific Art Assessment Network (SAAN), which can effectively exploit the style features and the generic aesthetic features of the given artwork. Our model consists of three modules: 1) **Generic Aesthetic Feature Extraction Branch:** inspired by the studies [27, 31], we adopt a self-supervised learning scheme to train a Generic Aesthetic Branch to extract aesthetics-aware features. The self-supervised scheme is based on the correlation between the aesthetic quality of the images and degradation editing operations. This essentially provides data augmentation such that the model can better learn the quality of different artworks. 2) **Style-specific Aesthetic Feature Extraction Branch:** observing that the style of the artwork is critical when assessing its aesthetic value and different styles need to extract different style-related aesthetic features, we propose a Style-specific Aesthetic Branch to incorporate style information into aesthetic features and extract style-specific aesthetic features via adaptive instance normalization [14]. 3) **Spatial Information Fusion:** we also add a non-local block [35] into the proposed method to fuse spatial information into the extracted aesthetic features.

The main contributions of our work are three-fold:

- We address the problem of artistic image aesthetics assessment, and introduce a new large-scale dataset BAID consisting of 60,337 artworks annotated with more than 360,000 votes to facilitate research in this direction.
- We propose a style-specific artistic image assessment network called SAAN, which combines style-specific and generic aesthetic features to evaluate artworks.
- We evaluate the state-of-the-art IAA approaches and our proposed method on the proposed BAID dataset. Our model achieves promising results on all the metrics, which clearly demonstrates the validity of our model.

Table 1. Summary of IAA/AIAA datasets and our proposed BAID dataset. BAID provides a significantly larger number of artistic images and has user subjective votes.

Dataset	Number of images	Number of artistic images
DP Challenge [4]	16,509	–
Photo.Net [16]	20,278	–
CUHK-PQ [22]	17,673	–
AVA [25]	255,530	–
AADB [17]	10,000	–
FLICKR-AES [28]	40,000	–
PARA [37]	31,220	–
TAD66K [12]	66,327	1,200
JenAesthetic [3]	1,628	1,628
VAPS [8]	999	999
BAID (Ours)	60,337	**60,337**

2. Related Work

Image Aesthetic Assessment Datasets. The Photo.net dataset [16] and the DPChallenge dataset [4] are the earliest attempts to construct public image databases for IAA. The Chinese University of Hong Kong-Photo Quality (CUHK-PQ) dataset is introduced in [22], which is the first dataset organized by topics. The AVA dataset [25] consists of approximately 255,000 images derived from DPChallenge.com with aesthetic annotations. Additionally, the AVA dataset contains photographic-style attributes and category attributes for a subset of images. Kong *et al*. [17] provided a new dataset called the Aesthetics and Attributes Database (AADB), which includes individual ratings of aesthetics and attributes of multiple images. Ren *et al*. [28] and Yang *et al*. [37] constructed FLICKR-AES and PARA respectively for personalized image aesthetic assessment. He *et al*. [12] introduced a theme-oriented dataset TAD66K which includes 47 themes and a unique criterion for each specific theme.

Although the above datasets have provided a solid foundation for IAA methods, they rarely include art images and consider different evaluation criteria for photos and artworks. As for existing AIAA datasets, neither of the public datasets Jenaesthetics [3] (1,628 art images) or VAPS (Vienna Art Picture System) [8] (999 paintings) is large enough to meet the requirements of deep learning methods.

In contrast, we construct the BAID dataset, which is, to our knowledge, the largest AIAA dataset made up entirely of artistic images (60,337 in total) and densely annotated

with scores (more than 360,000 votes). The comparison of our BAID and the existing datasets is listed in Tab. 1.

Image Aesthetic Assessment Models. Early studies on IAA mainly focus on designing and extracting handcrafted features from images and mapping the features to annotated aesthetics labels [7, 24]. With the emergence of large-scale IAA datasets [17, 25], methods based on deep learning continue to develop. NIMA [34] utilized Earth Mover's Distance (EMD) loss to predict the distribution of aesthetic scores. MP_{ada} [32] adopts an attention-based mechanism to dynamically adjust the weights of each patch during the training process to improve learning efficiency. Hosu *et al*. [13] propose the first AIAA method that efficiently supports full resolution images as an input, and can be trained on variable input sizes. [23] uses a saliency detection model to extract some more representative image patches, which are then fed into the network to extract features. She *et al*. [30] present a Hierarchical Layout-Aware Graph Convolutional Network (HLA-GCN) to capture layout information. TANet [12] can adaptively learn the rules for predicting aesthetics according to a recognized theme.

There are relatively few AIAA methods, where earlier traditional methods [2, 10, 18] design handcrafted features and train Support Vector Machine (SVM) for classification. Recently Zhang *et al*. [40] developed a deep multi-view parallel convolutional neural network (DMVCNN) to learn aesthetic features for Chinese ink paintings. In general, AIAA methods have not been adequately studied.

Different from the above works, we argue that different art styles need to extract different style-related aesthetic features, and combine both style-specific and generic aesthetic features to evaluate artworks.

3. Boldbrush Artistic Image Dataset

In this section, we discuss the data collection and the generation of scores of the proposed BAID dataset.

3.1. Data Collection

Constructing an artistic image dataset with score annotations is arduous. Most online art communities and professional artistic websites do not have public channels to score for artworks since the aesthetics of artworks are quite subjective and the scoring format is somewhat disrespectful to the artists, which has led, to some extent, to the inadequacy of the existing dataset.

We chose to use the website Boldbrush[1] as the source of data. Boldbrush hosts a monthly artwork contest where certified artists can upload their works and receive public votes from online users. Users can click into the detail page of the artwork and vote for it if they like the artwork, which means that the more votes, the greater the number of people consider the artwork pleasing and good-looking. Note that the users can vote for as many artworks as they like, and their individual votes are not ranked.

The benefits of our choice are as follows:

- The competition does not limit the subject matter, style or medium used to create the work, thus the website contains artworks with various art styles and contents.
- Every time a voter wants to place a vote, he/she will receive a verification email to confirm the vote. Moreover, the website performs email address check to prevent users from voting for the same work more than once. Thus, the votes will not suffer from malicious vote fraud and are more reliable than the 'favourite' annotations on other art communities like Flickr[2].
- Boldbrush and the FASO organization have a high profile in the art world. They have been holding such contests since July 2010. The voters are largely made up of artists and art collectors, so the results have a high degree of credibility and authority.

A total of 60,408 images and the corresponding annotations were collected, and 60,337 images are valid and available after removing corrupted data. Note that we exclude images with 0 vote since they are not included in the popular entries of BoldBrush.

3.2. Score Generation

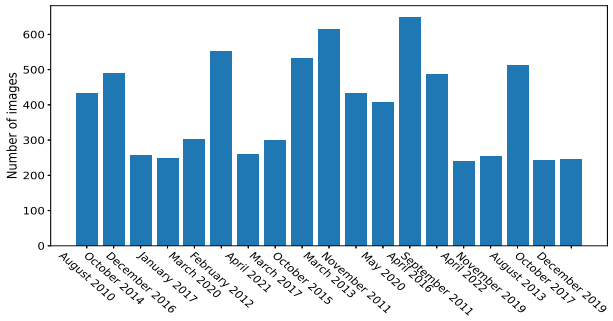
Unlike the existing IAA datasets where the score distribution is used to calculate the mean opinion score (MOS), we convert the number of votes to the scores of images in BAID. Simply put, images with a higher number of votes are considered to have a higher aesthetic value. Following the common practice, we choose to scale the number of votes into the [0, 10] score range, where 0 means the worst and 10 means the best. To elaborate the way we used to generate the scores, two characteristics of the contests' results need to be described:

- The number of votes received by entries in a month varies greatly. The margin between the highest number of votes and the lowest over a month can exceed 200.
- The images in the proposed dataset are all created by artists with a certain level of skills, thus the overall aesthetic quality is relatively high.

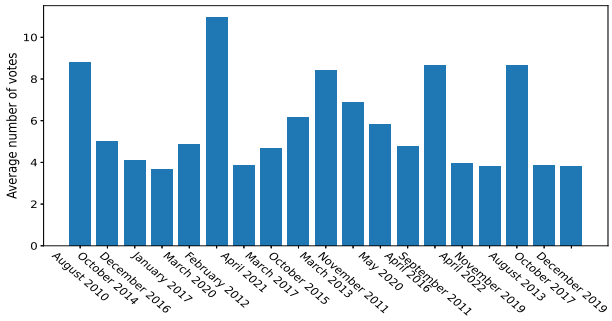
The distribution of the number of votes is shown in Fig. 2c. Since the margin between the highest number of votes and the lowest are too large to show in a figure (as described in Sec. 3.2), we choose images with 1 to 15 votes to demonstrate the overall distribution. Based on the two observations above, using linear mapping from votes to scores is not reasonable since it will make entries with low vote counts receive too low a score. After multiple attempts, we

[1] https://faso.com/boldbrush/popular

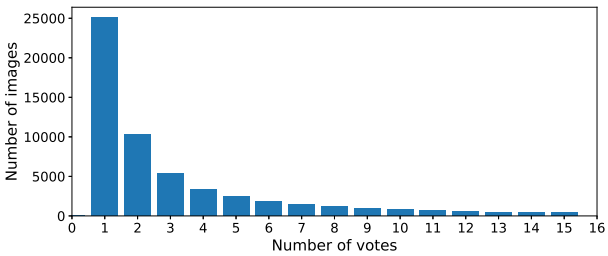
[2] https://www.flickr.com/



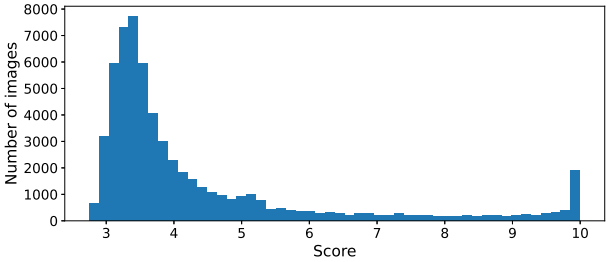
(a) The number of images collected in 20 randomly selected months.



(b) The average number of votes in 20 randomly selected months.



(c) The vote distribution of BAID.



(d) The score distribution of BAID.

Figure 2. Statistics of the proposed BAID.

Figure 3. Samples from BAID with generated scores (the number below is the aesthetic score of the image). Low score artworks are marked in red border and high score artworks are marked in green border.

adopt a sigmoid-like way to generate the scores. Specifically, given an image with the number of votes v_i and the entry month m_i, the score s_i is calculated using Eq. (1):

$$x_i = \frac{\bar{v}_{m_i} - v_i}{\bar{v}_{m_i}}, \qquad s_i = 10 \times \frac{1}{1 + e^{x_i}}, \tag{1}$$

where $\bar{v}_{m_i}$ is the average number of votes of month m_i. The final score distribution of the BAID is shown in Fig. 2d. Note that the original vote distribution is imbalanced and does not resemble a Gaussian distribution like most IAA datasets [4, 16, 17, 22, 25]: the number of images with low vote counts accounts for a large portion of the proposed dataset. While converting the number of votes to scores, we retain the characteristics of the original distribution due to the high credibility of the data source.

3.3. Further Analysis

To better demonstrate the data source (BoldBrush) and to support our selected score generating method, we randomly choose 20 months of data from BAID for illustration. Fig. 2a and Fig. 2b show the number of images and the average number of votes of the selected 20 months respectively. Fig. 2a demonstrates that the data sources are well balanced with the number of entries available each month being above 200, and the gap between months is not too large. Fig. 2b indicates that the average number of votes received by each month's entries varies, which means that the average aesthetic quality of the entries and the preference of voters may be different in each month of the contest. Thus, using a fixed threshold to generate binary labels or scores is not reasonable. Here we make use of the average number of votes received by the works in the month for normalization as described in Sec. 3.2.

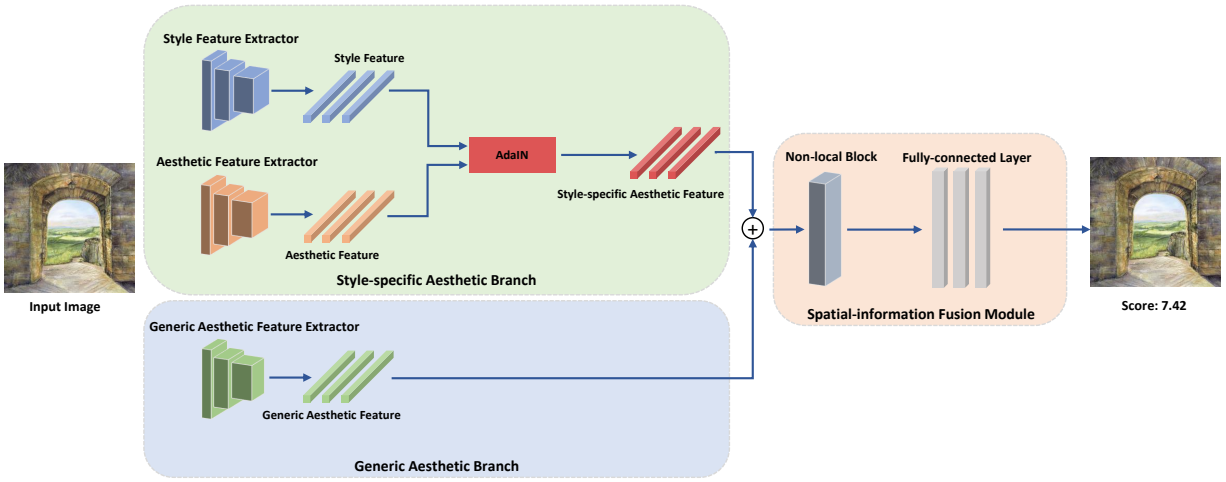


Figure 4. Overall architecture of the proposed SAAN. SAAN consists of three modules: 1) a style-specific branch to extract style-specific features; 2) a generic aesthetic branch to extract generic aesthetic features; and 3) a spatial information fusion module that fuses the spatial information using a non-local block and considers the composition of artwork during the assessment. See Sec. 4 for further details.

Since we do not have detailed information about the voters, we conducted an MOS (Mean Opinion Score) test to further validate our designed function Eq. (1). We sampled 100 artworks uniformly across the range of scores from the proposed BAID, and asked 10 college students majoring in art and design to score these samples. Results of the MOS test and more discussion of the score-generating function are given in Section 2 of the supplementary material.

4. Style-specific Art Assessment Network

In this section, we introduce our proposed approach Style-specific Art Assessment Network (SAAN), which uses style-specific and generic aesthetic features to evaluate artistic images. SAAN consists of three modules: 1) the Style-specific Aesthetic Branch (SAB) extracts style-specific aesthetic features (Sec. 4.1); 2) the Generic Aesthetic Branch (GAB) extracts generic aesthetic features based on self-supervised learning (Sec. 4.2); and 3) the Spatial Information Fusion Module fuses the spatial information using a non-local block and incorporates the composition of artwork into the assessment (Sec. 4.4). To train deep models to work better for aesthetic evaluation, we pretain the network by applying different manipulations, and training the network to classify manipulations and recognize manipulation intensity. The overall architecture of SAAN is displayed in Fig. 4.

4.1. Style-specific Aesthetic Feature Extraction

Intuitively, let us consider an oil painting p_1 and a pencil sketch p_2. From the perspective of human perception, when evaluating p_1, we may take into account the use of color and the variation of brushstrokes. Instead, we may put more emphasis on the control of lines when we evaluate p_2. Thus, the objective of the style-specific aesthetic branch is to extract the aesthetic features of the given artwork appropriate to its artistic style.

Style representations have been heavily discussed and studied in the field of style transfer. However, to the best of our knowledge, none of the existing IAA methods has considered the integration of style information into the prediction model. We follow the mainstream style transfer approaches [14, 19, 20, 26] to use an ImageNet [5] pretrained VGG-19 [33] backbone F_{sty} to extract style features. To extract aesthetics-related features, we use a ResNet-50 [11] backbone F_{aes}, which is pretrained by the self-supervised scheme in Sec. 4.3. Given an image p, the style features f_{sty} and the aesthetic features f_{aes} are extracted by:

$$\begin{aligned} f_{sty} &= F_{sty}(p, \theta_{sty}) \\ f_{aes} &= F_{aes}(p, \theta_{aes}), \end{aligned} \tag{2}$$

where θ_{sty} and θ_{aes} are the parameters of F_{sty} and F_{aes} respectively.

Instead of concatenating the style and aesthetic features together as input to subsequent network structures, we add an AdaIN [14] layer to integrate style information in f_{sty} into the aesthetic feature f_{aes}. Given a content feature map x and a style feature map y, AdaIN encodes the content and style information in the feature space by aligning the channel-wise mean and variance of x to match those of y:

$$AdaIN(x, y) = \sigma(y) \cdot \frac{x - \mu(x)}{\sigma(x)} + \mu(y). \tag{3}$$

Here we take the advantage that Huang *et al.* [14] mentioned in their study: the output produced by AdaIN will

have the same high average activation for the specific style feature, while preserving the spatial structure of the image.

The final output of the SAB is a style-specific aesthetic feature f_{aes_s} calculated as follows:

$$f_{aes_s} = AdaIN(f_{aes}, f_{sty}), \tag{4}$$

i.e., the style-specific aesthetic feature f_{aes_s} is obtained by changing the aesthetic feature f_{aes} to incorporate the style information of f_{sty}.

4.2. Generic Aesthetic Feature Extraction

In addition to the style-specific aesthetic branch, we propose a generic aesthetic branch to extract the aesthetic features shared by common categories of artworks. Aesthetic attributes like the integrity of the salient component and the layout of the frame can be viewed as intrinsic requirements.

We use ResNet-50 as the backbone to extract generic aesthetic and apply a self-supervised scheme to pretrain the backbone. Simply put, the pretraining stage includes two pretext tasks, one is to classify the applied distortions and the other is to estimate the intensity of the applied distortions. See Sec. 4.3 for further details. Denote the backbone as F_{gen}, the output generic aesthetic feature f_{aes_g} of a given image p is obtained by:

$$f_{aes_g} = F_{gen}(p, \theta_{gen}), \tag{5}$$

where θ_{gen} is the parameters of F_{gen}.

4.3. Self-supervised Pretraining Scheme

Pfister *et al.* [27] argue that ImageNet-pretrained backbones are not well-suited to the IAA task. For instance, such classification model should be invariant to the image's brightness and thus prohibits taking the image's brightness into account when evaluating its aesthetics. Sheng *et al.* [31] state that a trained IAA model is able to distinguish fine-grained aesthetic differences caused by various image manipulations. Based on the observation that certain distortions applied to images will reduce their appeal, both works [27, 31] proposed a self-supervised scheme to pretrain the backbone of an IAA model.

Inspired by these works, we adopt a pretraining approach similar to the one proposed in [31] in our work which includes two aesthetics-aware pretext tasks: one to identify the type of the distortion applied to a given image; and the other to detect the intensity of the applied distortion. The whole pretraining pipeline is shown in Fig. 5.

Degradation editing operations. Based on the selection of manipulations in [27, 31], we carefully select a variety of image manipulation operations to reduce artistic appeal. We design operations with different parameters for generating artificial training instances, which are listed in Tab. 2.

There are two main differences between the operations and the parameters we choose and the ones used in [31]:

Table 2. The operation list used in the pretraining pipeline. Operations marked in red color are our newly added ones which are not included in the operation list in [31].

Manipulation	Parameter
Gaussian noise	0.2, 0.4, 0.8
Quantization	64, 32, 8
Gaussian Blur	0.4, 0.8, 2
Exposure	1.5, 2.0, 2.5
Rotation	45, -45
Cropping	3/4, 2/3, 1/2
Stylization	(50, 0.6), (50, 0.3), (50, 0.1)
Convex	1/8, 1/4, 1/2
PencilSketch	(100, 0.1, 0.02), (100, 0.4, 0.02), (100, 0.6, 0.02)
CutMix [38]	32, 64, 128
None	-

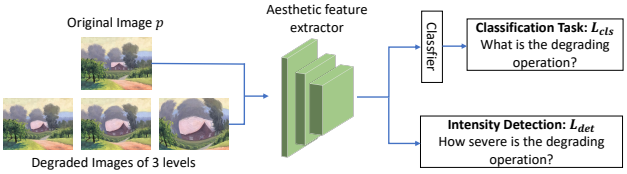


Figure 5. Pretraining pipeline. We first edit the original image using a distortion operation that reduces its appeal with three different levels. Then we train the aesthetic feature extractor with two pretext tasks: one to identify the type of the distortion, and the other to detect the intensity of the distortion.

- The operation list used in [31] ignores the distortion to some global aesthetic factors, *e.g.*, rule of thirds. We add operations that distort the layout and the composition of the original image (*e.g.* cropping, convex). We also add art-related distortions, *e.g.*, stylization which generates unwanted lines given a fine artwork.
- [31] only adopts two levels of distortions controlled by two sets of parameters. We carefully check the effect of the operations under different parameters and apply more subtle levels of distortion by using three sets of parameters (except for the rotation).

Distortion classification pretext task. This task is identical to the classification task proposed in [27, 31]. Denote an image patch as p, the loss term of the classification task $L_{cls}(p, t)$ reinforces the model to recognize which operation t has been applied to p:

$$\begin{aligned} L_{cls}(p, \theta_t) &= -\log(P_t(p; W)) \\ P_t(p; W) &= P(\hat{t} = t | m(p, \theta_t); W) \end{aligned} \tag{6}$$

where $m(p, \theta_t)$ is the manipulated output patch given the image patch p by the parameters θ_t, and $P(\hat{t} = t|m(p, \theta_t); W)$ is the probability predicted by our model W that p has undergone a degradation operation of type $\hat{t}$ that matches ground truth operation t. Note that for this task,

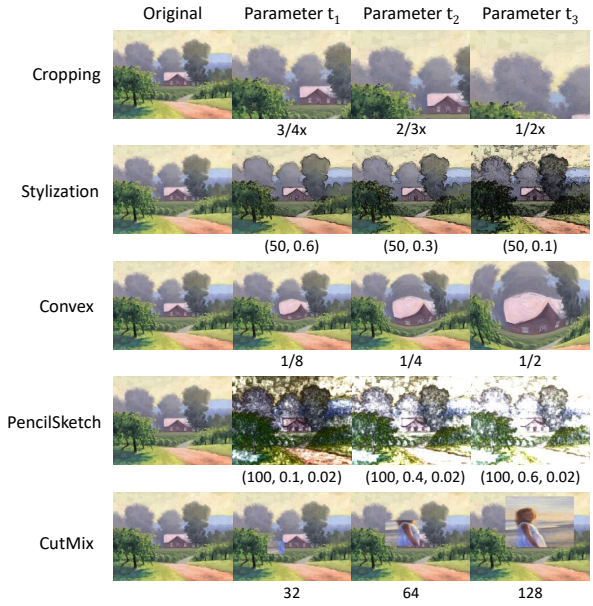


Figure 6. Visualization of the editing effects of the newly added operations in Tab. 2.

operations of different parameters are viewed as different operations.

Intensity detection pretext task. To detect the intensity of the distortion, Sheng *et al*. [31] proposed a triplet loss L_{trp}, which enforces a smaller distance between original patch and a slightly distorted patch, and a larger distance between original patch and a highly distorted patch. Given operation t and two control parameters θ_{t_1} and θ_{t_2}, L_{trp} is calculated using Eq. (7):

$$\begin{aligned} D(p,\theta_t) &= ||h(p,W) - h(m(p,\theta_t),W)||_2^2 \\ L_{trp}(p,\theta_{t_1},\theta_{t_2}) &= max\{0, 1 + D(p,\theta_{t_1}) - D(p,\theta_{t_2})\} \end{aligned} \tag{7}$$

where $h(p,W)$ is the L2-normalized feature of patch p extracted from the model with parameters W, and $D(p,\theta_t)$ works out the squared difference between the normalized features before and after applying the manipulation.

However, we find that two levels of distortion will make the L_{trp} hard to converge in our experiment. Meanwhile, the difference between the downgrading effect of θ_{t_1} and θ_{t_2} are sometimes too large, *e.g.*, Gaussion noise with $\theta_{t_1} = 0.2$ and $\theta_{t_2} = 0.8$ is not a smooth intensity transition. Thus, we apply **three levels of distortion** θ_{t_1}, θ_{t_2}, θ_{t_3} and introduce L_{det} as:

$$L_{det} = L_{trp}(p,\theta_{t_1},\theta_{t_2}) + L_{trp}(p,\theta_{t_2},\theta_{t_3}) - 1 \tag{8}$$

The overall loss of the pretraining pipeline is:

$$L = L_{cls} + \lambda L_{det}, \tag{9}$$

where λ is used to balance the two terms.

4.4. Spatial Information Fusion

Previous works [13, 30] have demonstrated that the layout of the given image is critical when predicting its aesthetic score. In this work, we add a non-local block [35] before the Multi-Layer Perception (MLP) to fuse the spatial information and implicitly detect the composition of the artwork. Specifically, given the extracted features f_{aes_s} and f_{aes_g} from the two branches mentioned above, the features are passed through one non-local block F_{nlb} after proper resizing and concatenation:

$$f_{out} = F_{nlb}(f_{aes_s} \oplus f_{aes_g}, \theta_{nlb}), \tag{10}$$

where $\oplus$ denotes the concatenate operation, and θ_{nlb} is the parameters of the non-local block.

Finally, an MLP L is used to output the predicted score s_{pred} of the input image p:

$$s_{pred} = L(f_{out}, \theta_L). \tag{11}$$

5. Experiments

In this section, we first describe the experiment settings used in the pretraining pipeline (introduced in Sec. 4.3), then we elaborate the training and evaluation of the proposed SAAN and state-of-the-art IAA methods on BAID, and conduct ablation studies to validate the effectiveness of each module.

5.1. Experimental Setup

Pretraining Settings. Instead of using a subset of ImageNet for pretraining [31], we directly use our BAID as the pretraining dataset to meet the requirements mentioned in [27]. We adopt the pretraining pipeline to train the ResNet-50 aesthetic feature extractor. For each image in BAID, we randomly choose three manipulation operations in Tab. 2 to edit it. We apply the Adam optimizer using a batch size of 64, with the weight decay of $5e-4$. We begin with a learning rate of $1e-3$, dropped it by a factor of 0.1 after every 10 epochs. Following the settings in [31], we activate L_{det} with $\lambda = 0.1$ after the first 30 epochs.

Training Settings. After pretraining, we then train the overall pipeline using the mean squared error (MSE) loss between predicted and ground truth aesthetic scores. Previous experiments on IAA [17,25] have reported that inappropriate data augmentation during training will degrade the performance at test time. Therefore, in the training stage, we directly resize the original image to 224×224 to avoid cropping which may decrease the aesthetic quality. We apply the Adam optimizer using a batch size of 64. We begin with a learning rate of $1e-5$, dropped it by a factor of 0.1 every 10 epochs for the first 40 epochs. Following the settings in [14], we freeze the VGG backbone in the style-specific aesthetic branch, and further freeze the ResNet-50

Table 3. Comparison with state-of-the-art open-source IAA methods on BAID.

Methods	#Params	SRCC ↑	PCC ↑	Accuracy ↑
NIMA [34]	63.61M	0.393	0.382	71.01%
MP_{ada} [32]	63.37M	0.437	0.425	74.33%
MLSP [13]	73.97M	0.441	0.430	74.92%
BIAA [41]	97.49M	0.389	0.376	71.61%
TANet [12]	57.87M	0.453	0.437	75.45%
Ours	64.44M	**0.473**	**0.467**	**76.80%**

Table 4. Ablation study results on the BAID.

Method	SRCC ↑	PCC ↑	Accuracy ↑
w/o style-specific branch	0.425	0.411	73.22%
w/o generic aesthetic branch	0.439	0.426	74.60%
w/o new editing operations	0.460	0.445	76.14%
w/o 3-level manipulation	0.462	0.448	76.19%
w/o spatial information fusion	0.459	0.440	76.14%
Ours	**0.473**	**0.467**	**76.80%**

backbone in the generic aesthetic branch to avoid overfitting into a certain category of artistic style. We randomly split the 60,337 images in BAID into 53,937:6,400 for training and testing respectively.

Evaluation Metrics. Typically, IAA methods are evaluated on regression and classification tasks. To evaluate the regression performance, we adopt two popular evaluation metrics: 1) Spearman's rank correlation coefficient (SRCC) S and 2) Pearson correlation coefficient (PCC) P. We also convert the predicted and ground-truth scores to binary-class labels (attractive & unattractive art) using a threshold of 5 (midpoint from 0 to 10) and calculate the accuracy.

5.2. Performance Comparison

We compare our method with five state-of-the-art open-source IAA methods on our BAID dataset, including NIMA [34], MP_{ada} [32], MLSP [13], BIAA [41] and TANet [12]. Note that most IAA methods are trained using EMD loss, which requires ground truth score distributions rather than only mean scores for training. Therefore, we modify the code provided by the researchers and make them trainable on BAID, which accounts for the reason that we only compare SAAN with open-source methods.

Tab. 3 shows the performance of the IAA methods and our SAAN on BAID. Compared with these methods, our SAAN model achieves the best performance on all metrics. This suggests that understanding the style of the artistic image assists in perceiving the aesthetics of the image, especially when there are a wide variety of styles that may have different evaluation criteria. For more performance evaluation, please refer to Section 3 of the supplementary material.

5.3. Ablation Study

Tab. 4 shows the ablation study results. (1) We first examine the effectiveness of the style-specific branch and the generic branch. All three metrics drop drastically when SAB is removed, where SRCC drops from 0.473 to 0.425, PCC drops from 0.467 to 0.411 and Accuracy drops from 76.80% to 73.22%. After removing the generic aesthetic branch, SRCC drops from 0.473 to 0.439, PCC drops from 0.467 to 0.426 and Accuracy drops from 76.80% to 74.60%. The disparity indicates that incorporating the style information with the generic information is of great help when evaluating artworks. (2) We then compare our new operation list (Tab. 2) and the one proposed in [31]. SAAN pretrained using our proposed list gives better results, which shows that adding global and art-related manipulations makes the model fit better in the artistic field. (3) We further analyze the efficacy of adopting 3 levels of distortions by deleting a set of parameters during pretraining, *i.e.*, using 2 levels. Results demonstrate that a more fine-grained intensity setting benefits the model to learn aesthetics-related features. (4) Finally, we remove the spatial information fusion module of the SAAN framework and the results demonstrate that fusing the spatial information enhances the performance of AIAA models.

6. Conclusions

This paper addresses the challenging task of artistic image aesthetic assessment (AIAA). To achieve this goal, we create a large-scale dataset BAID, which is constructed completely from artworks, including 60,337 artworks annotated with more than 360,000 votes. BAID is, to our knowledge, the largest artistic image aesthetic assessment dataset, and far exceeds existing IAA and AIAA datasets in quantity and quality of artworks. We further set up a complete benchmark and develop a baseline model called SAAN, which introduces adaptive perception to extract style-specific aesthetic features and achieves state-of-the-art performance on the proposed dataset. We hope our contributions will motivate the community to rethink AIAA and stimulate research with a broader perspective.

Acknowledgements. This work was supported by National Natural Science Foundation of China (72192821, 61972157, 62272447), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Science and Technology Commission (21511101200), Shanghai Sailing Program (22YF1420300, 23YF1410500), CCF-Tencent Open Research Fund (RAGR20220121) and Young Elite Scientists Sponsorship Program by CAST (2022QNRC001).

References

[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. ArtEmis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579, 2021. 2

[2] Seyed Ali Amirshahi and Joachim Denzler. Judging aesthetic quality in paintings based on artistic inspired color features. In *Proceedings of International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, 2017. 3

[3] Seyed Ali Amirshahi, Gregor Uwe Hayn-Leichsenring, Joachim Denzler, and Christoph Redies. JenAesthetics subjective dataset: analyzing paintings by subjective scores. In *Proceedings of the European Conference on Computer Vision*, pages 3–19. Springer, 2014. 2

[4] Ritendra Datta, Jia Li, and James Z Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *Proceedings of the IEEE International Conference on Image Processing*, pages 105–108. IEEE, 2008. 2, 4

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 5

[6] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017. 1

[7] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1657–1664. IEEE, 2011. 3

[8] Anna Fekete, Matthew Pelowski, Eva Specker, David Brieber, Raphael Rosenberg, and Helmut Leder. The Vienna Art Picture System (VAPS): A data set of 999 paintings and subjective ratings for art and aesthetics research. *Psychology of Aesthetics, Creativity, and the Arts*, 2022. 2

[9] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 1

[10] Xiaoying Guo, Takio Kurita, Chie Muraki Asano, and Akira Asano. Visual complexity assessment of painting images. In *Proceedings of IEEE International Conference on Image Processing*, pages 388–392, 2013. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[12] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 942–948. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track. 2, 3, 8

[13] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9375–9383, 2019. 1, 3, 7, 8

[14] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 1, 2, 5, 7

[15] Zhewei Huang, Wen Heng, and Shuchang Zhou. Learning to paint with model-based deep reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8709–8718, 2019. 1

[16] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5):94–115, 2011. 2, 4

[17] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *Proceedings of the European Conference on Computer Vision*, pages 662–679. Springer, 2016. 1, 2, 3, 4, 7

[18] Congcong Li and Tsuhan Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):236–252, 2009. 3

[19] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in Neural Information Processing Systems*, 30, 2017. 1, 5

[20] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. AdaAttN: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6649–6658, 2021. 1, 5

[21] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 17(11):2021–2034, 2015. 1

[22] Wei Luo, Xiaogang Wang, and Xiaoou Tang. Content-based photo quality assessment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2206–2213. IEEE, 2011. 1, 2, 4

[23] Shuang Ma, Jing Liu, and Chang Wen Chen. A-Lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4535–4544, 2017. 1, 3

[24] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1784–1791. IEEE, 2011. 3

[25] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *Proceedings of the IEEE Conference on Computer Vision*

and Pattern Recognition, pages 2408–2415. IEEE, 2012. 1, 2, 3, 4, 7

[26] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5880–5888, 2019. 1, 5

[27] Jan Pfister, Konstantin Kobs, and Andreas Hotho. Self-supervised multi-task pretraining improves image aesthetic assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 816–825, June 2021. 2, 6, 7

[28] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran. Personalized image aesthetics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 638–647, 2017. 2

[29] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015. 2

[30] Dongyu She, Yu-Kun Lai, Gaoxiong Yi, and Kun Xu. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8475–8484, 2021. 1, 3, 7

[31] Kekai Sheng, Weiming Dong, Menglei Chai, Guohui Wang, Peng Zhou, Feiyue Huang, Bao-Gang Hu, Rongrong Ji, and Chongyang Ma. Revisiting image aesthetic assessment via self-supervised feature learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5709–5716, 2020. 2, 6, 7, 8

[32] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. Attention-based multi-patch aggregation for image aesthetic assessment. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 879–886, 2018. 1, 3, 8

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[34] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. 1, 3, 8

[35] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 2, 7

[36] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. BAM! the Behance artistic media dataset for recognition beyond photography. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1202–1211, 2017. 2

[37] Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19861–19869, 2022. 2

[38] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 6

[39] Cunjun Zhang, Kehua Lei, Jia Jia, Yihui Ma, and Zhiyuan Hu. AI painting: an aesthetic painting generation system. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 1231–1233, 2018. 1

[40] Jiajing Zhang, Yongwei Miao, Junsong Zhang, and Jinhui Yu. Inkthetics: A comprehensive computational model for aesthetic evaluation of chinese ink paintings. *IEEE Access*, 8:225857–225871, 2020. 3

[41] Hancheng Zhu, Leida Li, Jinjian Wu, Sicheng Zhao, Guiguang Ding, and Guangming Shi. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *IEEE Transactions on Cybernetics*, 2020. 8