Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks

Shuai He , Yongchang Zhang , Rui Xie , Dongxiang Jiang , Anlong Ming*
Beijing University of Posts and Telecommunications
{hs19951021, zhangyongchang, mibxr, jiangdx, mal}@bupt.edu.cn

Abstract

Challenges in image aesthetics assessment (IAA) arise from that images of different themes correspond to different evaluation criteria, and learning aesthetics directly from images while ignoring the impact of theme variations on human visual perception inhibits the further development of IAA; however, existing IAA datasets and models overlook this problem. To address this issue, we show that a theme-oriented dataset and model design are effective for IAA. Specifically, 1) we elaborately build a novel dataset, called TAD66K, that contains 66K images covering 47 popular themes, and each image is densely annotated by more than 1200 people with dedicated theme evaluation criteria. 2) We develop a baseline model, TANet, which can effectively extract theme information and adaptively establish perception rules to evaluate images with different themes. 3) We develop a large-scale benchmark (the most comprehensive thus far) by comparing 17 methods with TANet on three representative datasets: AVA, FLICKR-AES and the proposed TAD66K, TANet achieves state-of-the-art performance on all three datasets. Our work offers the community an opportunity to explore more challenging directions; the code, dataset and supplementary material are available at https://github.com/woshidandan/TANet.

1 Introduction

Can you tell which image in Fig. 1 is more beautiful? It is difficult to assess and compare the aesthetic perception of images with different themes because they correspond to different evaluation criteria, e.g., large and strong trees versus small and delicate flowers, all of which may be beautiful. Classic photography [Barnbaum, 2017] has shown that image aesthetics are strongly related to the theme; to understand the aesthetics of an image, photographers first pay attention to the theme of the image and then apply different evaluation criteria for images with different themes (image→theme→aesthetics).

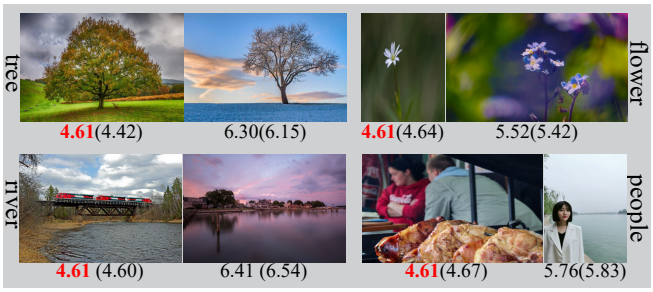

Figure 1: Examples of images and annotations in the proposed TAD66K dataset with the ground truth (and predicted) scores shown below each image. Even with the same scores, aesthetics of different themes correspond to different criteria and cannot be compared directly.

Utilizing several existing datasets (Table 1), previous studies directly learned aesthetics from images annotated with **aesthetic scores**. However, images are composed explicitly of natural signals; in other words, the elements based on pixels cannot contain any abstract or thematic information. The aesthetic scores given by the human subconscious are related to themes, but existing methods do not consider or take advantage of this. Thus, existing methods have attempted to build image→aesthetic mappings, which violate the process of human visual perception.

This mapping approach worsens image aesthetics assessment (IAA) in two respects: 1) The annotations of existing datasets do not consider that different themes have different scoring criteria; furthermore, all images are mixed together and scored without differentiation, which introduces considerable noise and error into the ground truth. 2) Even using the most advanced methods, learning directly from low-level pixels while using noisy ground truth as the supervision information makes it difficult to effectively perceive aesthetic information, and these methods are not powerful enough to understand the aesthetics, which causes attention dispersion (Fig. 2 and Fig. 3). These limitations inhibit the development of IAA, which is not yet fully understood.

To comprehensively investigate this topic, we provide two

*Contact Author. This work was supported by the national key R & D program intergovernmental international science and technology innovation cooperation project (2021YFE0101600).

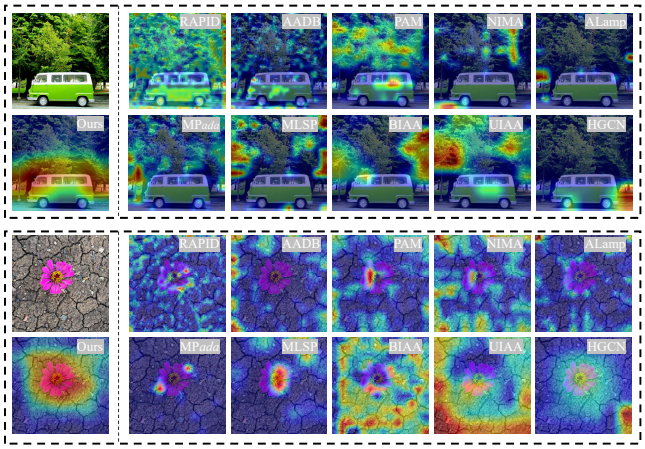

Figure 2: Saliency map comparisons of the 10 SOTA models. Our TANet captures the high aesthetic area successfully.

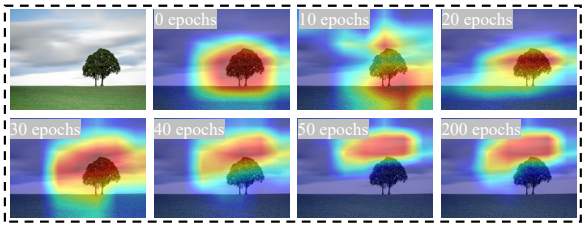

Figure 3: Previous work pretrained a model on ImageNet to understand basic information; however, this understanding ability is lost in the training process of IAA tasks. We call this phenomenon "attention dispersion".

Dataset	Image	Rating	Theme
DP Challenge [Datta *et al.*, 2008]	16,509	100	N/A
Photo.Net [Joshi *et al.*, 2011]	20,278	500	N/A
CUHK-PQ [Luo *et al.*, 2011]	17,673	10	7
AVA [Murray *et al.*, 2012]	**255,530**	250	N/A
AADB [Kong *et al.*, 2016]	10,000	5	N/A
PCCD [Chang *et al.*, 2017]	4,235	7	N/A
FLICKR-AES [Ren *et al.*, 2017]	40,000	210	N/A
DPC-Captions [Jin *et al.*, 2019]	154,384	6	N/A
SPAQ [Fang *et al.*, 2020]	11,125	600	9
TAD66K (Ours)	66,327	**12,00**	**47**

Table 1: Summary of previous IAA datasets and TAD66K. TAD66K provides much richer ratings and themes.

contributions. First, we build a large-scale dataset called the **T**heme and **A**esthetics **D**ataset with **66K** images (***TAD66K***), which is specifically designed for IAA. Specifically, 1) it is a ***theme-oriented*** dataset containing 66K images covering 47 popular themes. All images were carefully selected by hand based on the theme. 2) In addition to common aesthetic criteria, we provide 47 criteria for the 47 themes. Images of each theme are annotated independently, and each image contains at least 1200 effective annotations (so far the richest annotations). These high-quality annotations could help to provide deeper insight into the performance of models.

Second, we propose a baseline model, called the **T**heme and **A**esthetics **Net**work (***TANet***), which can maintain a constant perception of aesthetics to effectively deal with the problem of attention dispersion. Moreover, TANet can adaptively learn the rules for predicting aesthetics according to a recognized theme. To further improve the perception of each theme, we propose an RGB-distribution-aware attention network (RGBNet) to help the network perceive the color distribution in the RGB space and solve the problems associated with the high complexity of standard attention. Additionally, using TAD66K and two existing datasets, namely, Aesthetic Visual Analysis (AVA) [Murray *et al.*, 2012] and Flickr Images with Aesthetics Annotation Dataset (FLICKR-AES) [Ren *et al.*, 2017], which are widely used to verify general and personalized aesthetic models, we offer a rigid evaluation of 17 state-of-the-art (SOTA) baselines (Table 2), making this work the most complete IAA benchmark thus far. Promising results are achieved on all of the above benchmarks, clearly demonstrating the effectiveness of our model.

2 Related Work

IAA Datasets. Over the past few years, several datasets have been constructed for IAA (Table 1). The datasets DP Challenge [Datta *et al.*, 2008] and Photo.Net [Joshi *et al.*, 2011] were adopted early on, but they contain relatively few images, and these images comprise coarse annotations. Murray *et al.* [Murray *et al.*, 2012] created AVA, a large-scale IAA dataset with nearly 255,000 images that has become one of the most popular IAA datasets. Subsequently, Kong *et al.* [Kong *et al.*, 2016] provided a novel dataset called the Aesthetics and Attributes Database (AADB) that includes the aesthetic and attribute scores of multiple images by individual users. Yu *et al.* [Chang *et al.*, 2017] and Jin *et al.* [Jin *et al.*, 2019] compiled datasets called the Photo Critique Captioning Dataset (PCCD) and DPC-Captions, respectively, to solve the aesthetic-related photo caption generation problem. Ren *et al.* [Ren *et al.*, 2017] created FLICKR-AES for personalized image aesthetics. CUHK-PQ [Luo *et al.*, 2011] is the first dataset to organize images by theme, but it contains only 7 themes. More recently, Fang *et al.* [Fang *et al.*, 2020] introduced the Smartphone Photography Attribute and Quality (SPAQ) dataset with 9 themes. Although the above datasets have advanced the field of IAA to various degrees, they fail to consider that the same evaluation criteria are not suitable for highly variable themes, which causes a label noise problem.

Unlike those of the existing datasets, the goal of TAD66K is to provide a more challenging and theme-oriented dataset. To the best of our knowledge, TAD66K is the largest densely annotated dataset (Table 1). Furthermore, TAD66K provides carefully selected themes, targeted evaluation criteria, and measures to prevent the long-tailed effect that occurs in the existing datasets; these features make TAD66K a solid and high-quality dataset.

IAA Models. The early models described in Table 2 focused mainly on extracting aesthetic information from im-

No.	Model	Pub.	Training Set	Basic	Code
1	RAPID [Lu *et al.*, 2014]	ACMMM	AVA	incorporate heterogeneous	Python&Lua
2	DMA [Lu *et al.*, 2015]	ICCV	AVA	multi-patch aggregation	N/A
3	MNA [Mai *et al.*, 2016]	CVPR	AVA	adaptive spatial pooling	N/A
4	AADB [Kong *et al.*, 2016]	ECCV	AADB+AVA	sampling strategy, rangking loss	MATLAB
5	PAM [Ren *et al.*, 2017]	ICCV	AES+CUR	residual-based, active learning	Caffe
6	ALamp [Ma *et al.*, 2017]	CVPR	AVA	layout-aware, multi-patch	Scipy
7	NIMA [Talebi and Milanfar, 2018]	TIP	AVA	predict distribution	Tensorflow
8	MP_{ada} [Sheng *et al.*, 2018]	ACMMM	AVA	attention, multi-patch	Tensorflow
9	CFAN [Wang *et al.*, 2018]	IJCAI	AVA	collaborative filterin	N/A
10	MLSP [Hosu *et al.*, 2019]	CVPR	AVA	staged training,multi-level features	Tensorflow
11	BIAA [Zhu *et al.*, 2020]	TCYB	AES+CUR+AADB	meta-learning, bilevel optimization	PyTorch
12	UIAA [Zeng *et al.*, 2019]	TIP	AVA+AADB	unified probabilistic formulation	MATLAB
13	AFDC [Chen *et al.*, 2020]	CVPR	AVA	fractional dilated kernel	N/A
14	PIAA [Li *et al.*, 2020]	TIP	AVA+AES	personality-assisted multi-task	N/A
15	UGIAA [Lv *et al.*, 2021]	TMM	AVA+AES	deep reinforcement learning	N/A
16	MUSIQ [Ke *et al.*, 2021]	ICCV	AVA	multi-scale representation	N/A
17	HGCN [She *et al.*, 2021]	CVPR	AVA+AADB	graph convolution networks	Jittor
18	TANet (Ours)	IJCAI	AVA+AES+TAD66K	attention, adaptive features	PyTorch

Table 2: Summary of 17 existing representative IAA models and the proposed TANet. **Training Set**: We count only datasets related to IAA. **Code**: N/A means that the official code is not available online; available code links are marked in red.

ages and mapping the visual features to annotated labels by training classifiers (aesthetically positive or negative) or regressors [Lu *et al.*, 2014]. Although these methods have achieved great success, recent evidence reveals that the direct prediction of aesthetic scores removes the diversity (distribution) of human opinions [Talebi and Milanfar, 2018]. Some researchers have noted this limitation and proposed using the Earth mover's distance (EMD) loss [Talebi and Milanfar, 2018] to train the score distribution task, while others have attempted to adapt a personalized aesthetics model for individual preferences [Ren *et al.*, 2017; Lv *et al.*, 2021].

However, previous works ignore the importance of theme information, when humans assess aesthetics, they explicitly or implicitly take into account the influence of the themes; additionally, relying only on the weak supervision information of IAA datasets is not sufficient to understand aesthetics (Fig. 2). In this work, we focus on finding theme-adaptive understanding methods and investigating how to best use them for IAA to yield a competitive model.

3 Proposed Dataset

The emergence of new datasets has led to rapid progress on various IAA tasks. In this context, our goals for developing TAD66K are to provide a new challenge and to spark novel ideas to rethink IAA tasks. We describe details as follows.

3.1 Image Collection

In terms of data collection, we took the following steps. First, we collected themes that are widely popular among humans. To ensure the richness of these themes, we determined the most uploaded themes on the Flickr website from 2008 to 2021. With the help of the t-distributed stochastic neighbor embedding (t-SNE) technique [Van der Maaten and Hinton, 2008], we manually grouped the themes into seven superthemes, namely, plants, animals, artifacts, colors, humans, landscapes and other, which were further divided into 47 subthemes, as shown in Fig. 4.

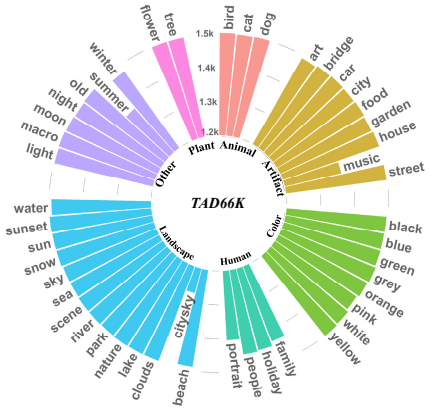

Figure 4: The histogram distribution of the TAD66K dataset.

Second, we reduced the long-tailed distributions of the annotated data by performing a preassessment during the data collection phase. Imbalanced data were ignored in previous works; for example, the AVA dataset, one of the largest aesthetic datasets currently available, contains annotated image aesthetic scores from 1 to 10, but we find that the majority score range (5∼6) has 2700 times more samples than the minority score range (9∼10), models trained on imbalanced data are biased toward majority examples. To alleviate the disadvantages of previous datasets, we preliminarily assessed the aesthetics of samples as good, general, or bad and tried to ensure that the samples are evenly divided among each of these three labels to maximize the diversity of the overall score.

3.2 Image Annotation

We set the annotation score of each image to range from 1 to 10, representing the range from the lowest aesthetics to the highest aesthetics. According to some theories of photography, different themes follow different evaluation criteria (see our **supplementary material**), so we labeled each theme separately. A common problem of subjective rating experiments is that the annotators often have no baseline (or reference) for assessing the measured effect. This often causes the subjective responses obtained before the annotators have seen adequate examples to be unreliable. To alleviate this, we provided anchor images for different rating scales as references and employed batch annotation (batch = 50). During the annotation process, the annotators needed to browse the anchor images and 50 images completely before evaluating the images. Finally, through manual cleaning, approximately 1200 opinions were collected on average for each image. We calculated the average value as the ground truth of the image.

4 Proposed Model

Overview. The general IAA model is described as follows:

$$p = F(x, \theta), \tag{1}$$

where F indicates the IAA model, p represents the predicted aesthetic score, and x represents the input image. However, it is difficult to use only single aesthetic scores as supervision information to extract perceptually consistent features. Nevertheless, humans visually perceive and evaluate aesthetics after the image theme is recognized. Furthermore, the weight parameters θ ignore how theme variations influence the method with which aesthetics are perceived. With the above problems in mind, the proposed TANet model has three essential components (Fig. 5), which we describe as follows.

4.1 Theme Understanding Network

To extract theme features via direct supervision, we use ResNet18 as the backbone S and train it on the scene database [Zhou *et al.*, 2017], achieving a 85.03% top-5 accuracy. The scene database is a repository of 10 million images annotated with 400+ unique theme semantic categories and environments, and almost the themes in TAD66K dataset and daily life are covered. It is worth noting that to prevent a gradual decline in the perception ability (Fig. 3) during IAA model training, we *freeze* the parameters at all times. We define the output of S as $S(x)$, which is divided into two flows for processing. One flow is sent to a parameter generator L^1 to adaptively generate θ_{theme} (weights and bias):

$$\theta_{theme} = L^1(S(x), \delta), \tag{2}$$

where δ represents the parameters of L^1. To deal with the noise in $S(x)$, the other flow is sent to the feature preprocessor L^2 to reduce the spatial redundancy in the latent representation. Finally, we use a linear layer L^3 to multiply the two flows to obtain the final output, and the whole process can be expressed as:

$$x_{theme} = L^3(L^2(S(x)), \theta_{theme}). \tag{3}$$

Thus, x_{theme} includes both the basic theme information and the rules governing how to perceive this information.

4.2 RGB-distribution-aware Attention Network

The second component extracts high-level color features from the RGB space to improve the understanding of the theme. The color distribution [O'Donovan *et al.*, 2011] constitutes important information in aesthetics and has a close relationship with the theme. In terms of capturing this distribution, the standard self-attention mechanism [Vaswani *et al.*, 2017] can coordinate the relationships between a token and all other tokens in natural language processing (NLP) tasks. However, for IAA tasks, the original information of the color distribution and similarity exists in low-level features closer to the original image, and a large feature map size means that there are more tokens; thus, global computations are of quadratic complexity [Liu *et al.*, 2021].

In our implementation, we apply two improvements to the standard self-attention mechanism. We first split the input into nonoverlapping patches, where each patch is represented by a center point and is set as the average of the raw pixel RGB values. For images of any size, the patch space is composed of $k \times k$ central points, which results in a linear and low computational complexity to the input size. Specifically, we set $k = 12$. Second, we extract only the relationship of patches without multiplying them by the input. Let f_{ab}, f_{uv} represent two centers; then, the output x_{rgb} of our attention can be described as:

$$x_{rgb} = \|_{l=1}^{N} \left(\text{Softmax} \left(L^4 \left(\frac{(Q^l f_{ab})^T (K^l f_{uv})}{\sqrt{d}} \right) \right) \right), \tag{4}$$

where $\|_{l=1}^{N}$ is the concatenation of RGB channels and Q^l, K^l and d are the query, key and dimension, respectively, generated from the input in the standard self-attention. After extracting this relationship, we do not multiply it by the value (V) but instead send it to the softmax layer after processor L^4 to reduce redundant information and obtain the final result x_{rgb}. As a result, the proposed attention mechanism endows the RGB space with a perceptually meaningful measure of the color similarity and distribution, as shown in Fig. 6.

4.3 Aesthetics Perceiving Network

The third component performs two functions. First, APNet directly extracts aesthetic features x_{aes} from the input x; we use MobileNetV2 as the backbone, and the output is processed by L^5. Second, three features are fused to predict an aesthetic score, and the output is processed by L^6. We describe the whole process as:

$$p = F_{aes}(x_{theme} \oplus x_{rgb} \oplus x_{aes}, \theta_{aes}), \tag{5}$$

where θ_{aes} represents all the parameters of F_{aes}. Please refer to the **supplementary material** for more details.

5 Experiments

5.1 Experimental Settings

Evaluation Metrics. To evaluate the performance, we adopt two popular evaluation metrics, namely, Spearman's rank correlation coefficient (**SRCC**) $\mathcal{S}$ [Talebi and Milanfar, 2018] and the linear correlation coefficient (**LCC**) $\mathcal{L}$ [Talebi and Milanfar, 2018]; the evaluation indicators for the AVA loss

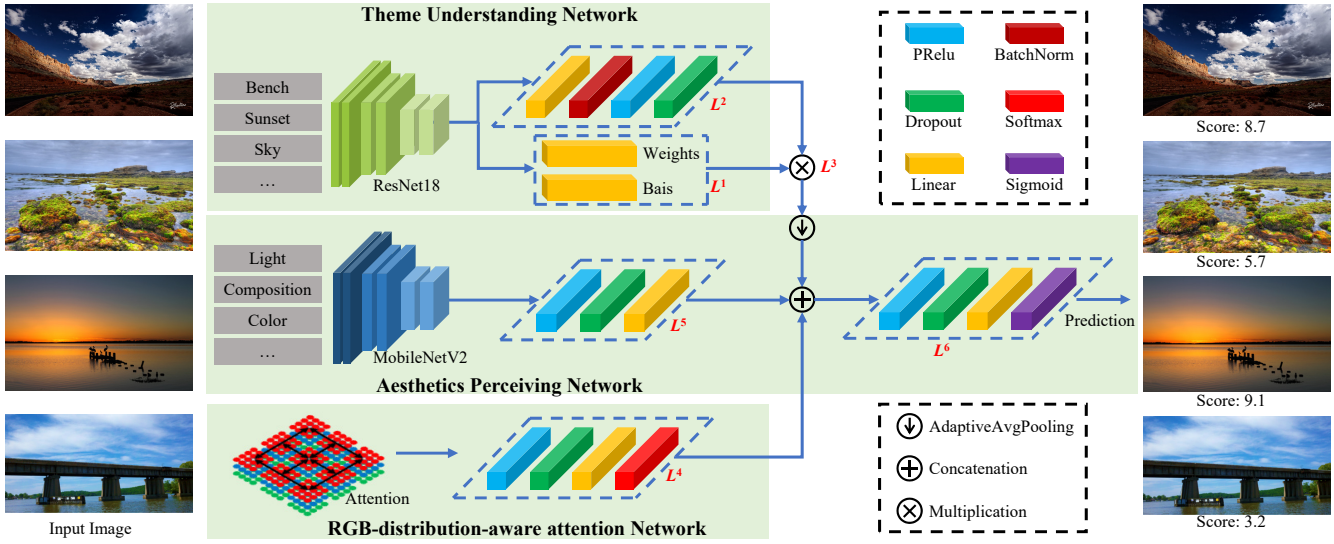

Figure 5: Overall architecture of the proposed TANet model. TANet consists of three components: TUNet is intended to efficiently learn theme information, RGBNet obtains the color similarity and distribution relationship, and APNet is introduced to fuse multiscale features to predict the aesthetic score. See section 4 for further details.

	Metric	Code Available 2014-2021										Code Not Available 2015-2021							Ours
		RAPID	AADB	PAM	NIMA	ALamp	MP_{ada}	MLSP	BIAA	UIAA	HGCN	DMA	MNA	CFAN	AFDC	PIAA	UGIAA	MUSIQ	
AVA	$\mathcal{S}\uparrow$	.447*	.558	.712*	.612	.666*	.727	.756	.651*	.719	.665	-	-	-	.648	.677	.692	.726	**.758**
	$\mathcal{L}\uparrow$	.453*	.580*	.715*	.636	.671*	.731	.757	.668*	.720	.687	-	-	-	.671	-	-	.738	**.765**
	$\mathcal{E}\downarrow$	-	-	-	.050	-	-	-	-	.065	**.043**	-	-	-	.044	.047	-	-	.047
	$\mathcal{R}\uparrow$	.628*	.722	.876*	.751	.807*	.875	.925	.853*	.890	.786	?/.754	?/.774	?/.810	.779	.809	.813	.726/?	**.940**
TAD66K	$\mathcal{S}\uparrow$	.314*	.379*	.422*	.390*	.411*	.466*	.490*	.417*	.433*	.486*	-	-	-	-	-	-	-	**.513**
	$\mathcal{L}\uparrow$	.332*	.400*	.440*	.405*	.422*	.480*	.508*	.431*	.441*	.493*	-	-	-	-	-	-	-	**.531**
	$\mathcal{M}\downarrow$	.022*	.021*	.020*	.021*	.019*	.022*	.019*	.020*	.021*	.020*	-	-	-	-	-	-	-	**.016**

Table 3: Comparison of 17 state-of-the-art IAA models on 2 datasets: AVA [Murray *et al.*, 2012] and the proposed TAD66K dataset. For some models with publicly available codes, we use the recommended parameter settings to obtain some metrics ('*') that are not available to complete the benchmarks. '-' or '?' mean the metric is not available in the paper. See subsection 5.1 for details regarding the metrics.

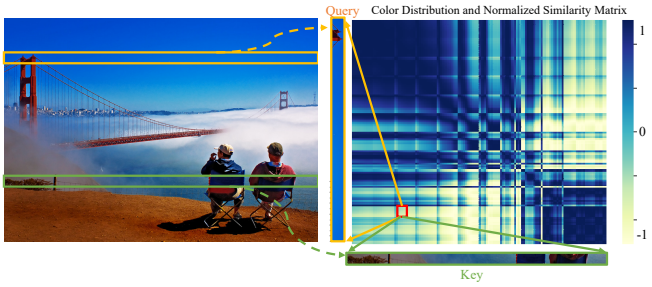

Figure 6: Visualization of the color distribution and similarity exists in original image that extracted by our attention module.

Model	10 images	100 images
PAM(att&con)	0.520±0.003	0.553±0.012
PIAA	0.543±0.003	0.639±0.011
UGIAA	0.559±0.002	0.660±0.013
BIAA	0.561±0.005	0.669±0.013
Ours	**0.609±0.005**	**0.717±0.011**

Table 4: Comparison results (SRCC) of the state-of-the-art models for personalized aesthetics assessment on the FLICKR-AES dataset [Ren *et al.*, 2017].

also include the **EMD** loss $\mathcal{E}$ [Talebi and Milanfar, 2018], while adopting the SRCC/accuracy **ratio** $\mathcal{R}$ [Hosu *et al.*, 2019] instead of the binary classification accuracy to better evaluate the generalized performance. For TAD66K, we adopt the mean squared error (**MSE**) $\mathcal{M}$ as the loss function.

Benchmark Models. We select 17 SOTA models on the AVA dataset that 1) are recently published and 2) have a representative pipeline and apply 10 models (whose code is available) to our TAD66K dataset. There are 4 other designs specifically for personalized IAA that we use to evaluate performance on the FLICKR-AES dataset.

Type	$\mathcal{S}\uparrow$	$\mathcal{L}\uparrow$	$\mathcal{M}\downarrow$	Speed ↑
APNet	0.440	0.457	0.020	**13.89 it/s**
(AP + TU)Net	0.483	0.513	0.017	10.03 it/s
(AP + RGB)Net	0.462	0.478	0.019	13.71 it/s
APNet + self-attention	-	-	-	out memory
(AP + RGB + TU)Net	**0.513**	**0.531**	**0.016**	10.01 it/s

Table 5: Ablation studies of TANet on TAD66K (single 2080Ti).

5.2 Performance Comparison

Performance on TAD66K. Compared with the 10 SOTA methods (Table 3), our TANet model achieves the best performance on all metrics. This suggests that understanding the theme of an image fundamentally assists in perceiving the aesthetics of the image, especially when there are a wide variety of themes that may have different evaluation criteria. Some examples are shown in Fig. 1. In addition, a real-time inference video is available in the **supplementary material**.

Performance on AVA. Based on the overall performance of the 17 representative methods reported in Table 3, TANet still achieves the top performance on the popular AVA dataset. This is because its specially designed TUNet and RGBNet can automatically learn general high-level theme features and color features, which are crucial for better capturing the general evaluation criteria for a general aesthetic assessment.

Performance on FLICKR-AES. Previous work has shown that people with different personalities may prefer images with specific themes [Li *et al.*, 2020; Ren *et al.*, 2017]. When our network understands the theme information, it is able to notably improve the overall personalized IAA performance. Table 4 shows that our model achieves the best SRCC of 0.717, surpassing the second-best results by +7.2% SRCC, which means that our model can use a small amount of data to learn more about individual users' aesthetic preferences.

5.3 Ablation Study

We first examine the effectiveness of TUNet (Table 5), which obtains an SRCC of +9.8% and an LCC of +12.3%, while the MSE loss decreases by 15.0%. However, TUNet needs to extract theme features, which reduces the training speed of the model to 3.68 it/s. Furthermore, when RGBNet is added, the performance achieves increases in the SRCC and LCC of +5.0% and +4.6%, respectively, and the MSE loss value is decreased by 5.0%. Moreover, the proposed attention mechanism in RGBNet model does not significantly reduce the speed of the network. We also conduct a comparison test with the standard self-attention and find that due to its excessive computational complexity, it cannot work properly in the same way, which shows that proposed architecture is more computationally efficient than the traditional architecture. Finally, we add all elements. The SRCC and PLCC values are further improved to their highest values, which are +16.6% (SRCC) and +16.2% (LCC), while the MSE loss value is decreased by 20.0%, which confirms that the proposed elements are effective in promoting the model's performance.

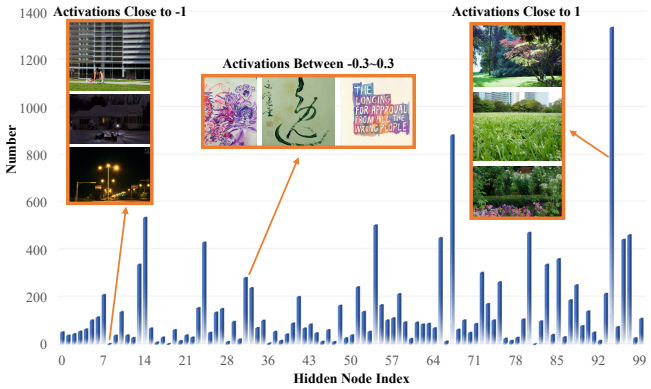

Figure 7: The adaptive parameters θ_{theme} include 100 hidden nodes (indexed from 0 to 99), and images with specific themes are activated by specific hidden nodes. If the activation of one node is maximally positive or minimally negative, then the count is increased by one. We verify TANet on test sets, and this figure indicates that our TANet adopts distinct criteria for evaluating different themes.

5.4 Model Interpretation

Adaptive Weights. To discern whether the model adaptively adjusts the weights of theme feature fusion, we extract the weights of θ_{theme} in Eq.(2), which includes 100 hidden nodes. We verify the trained TANet model on the test set of TAD66K, count the index of each image when it is activated in the maximum positive or minimum negative direction, and plot the results in Fig. 7. Interestingly, images with specific themes are activated by specific hidden nodes. e.g., TANet tends to assign lower negative weights (close to -1) to images with low aesthetics since these images usually have chaotic themes or are too dark. Among the most frequently visited nodes, images with harmonious colors and bright themes have more positive responses. Additionally, we find that some hidden nodes respond only to specific themes, such as images with the theme "art". This verifies that our model has learned to build the corresponding evaluation criteria.

Class Activation Maps. To visualize saliency maps, we fuse the 2-D feature maps of the last layers of the three elements in TANet, and we do the same for the other 10 SOTA methods with multiple branches. The attention of our model (Fig. 2) is able to cover more highly theme-correlated areas and is more in line with human perception. Furthermore, our model effectively solves the problem of attention dispersal.

6 Conclusions

This paper addresses the long-ignored influence of theme variation in IAA. To achieve this goal, we create a theme-oriented TAD66K dataset (containing 47 themes), build up a complete benchmark (including the top 17 SOTA models) and develop a baseline model called TANet. Compared with the existing datasets, TAD66K is more challenging and more densely annotated; moreover, the proposed TANet introduces adaptive perception methods to extract theme features and achieves SOTA performance on three representative datasets. We hope our contributions will motivate the community to rethink IAA and stimulate research with a broader perspective.

References

[Barnbaum, 2017] Bruce Barnbaum. *The Art of Photography: A Personal Approach to Artistic Expression*. Rocky Nook, Inc., 2017.

[Chang *et al.*, 2017] Kuang-Yu Chang, Kung-Hung Lu, and Chu-Song Chen. Aesthetic critiques generation for photos. In *ICCV*, pages 3514–3523, 2017.

[Chen *et al.*, 2020] Qiuyu Chen, Wei Zhang, Ning Zhou, Peng Lei, Yi Xu, Yu Zheng, and Jianping Fan. Adaptive fractional dilated convolution network for image aesthetics assessment. In *CVPR*, pages 14114–14123, 2020.

[Datta *et al.*, 2008] Ritendra Datta, Jia Li, and James Z Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *ICIP*. IEEE, 2008.

[Fang *et al.*, 2020] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *CVPR*, pages 3677–3686, 2020.

[Hosu *et al.*, 2019] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *CVPR*, 2019.

[Jin *et al.*, 2019] Xin Jin, Le Wu, Geng Zhao, Xiaodong Li, Xiaokun Zhang, Shiming Ge, Dongqing Zou, Bin Zhou, and Xinghui Zhou. Aesthetic attributes assessment of images. In *ACMMM*, pages 311–319, 2019.

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

[Ke *et al.*, 2021] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021.

[Kong *et al.*, 2016] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, pages 662–679. Springer, 2016.

[Li *et al.*, 2020] Leida Li, Hancheng Zhu, Sicheng Zhao, Guiguang Ding, and Weisi Lin. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. *IEEE TIP*, 29:3898–3910, 2020.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv:2103.14030*, 2021.

[Lu *et al.*, 2014] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z. Wang. RAPID: Rating pictorial aesthetics using deep learning. In *ACMMM*, pages 457–466, 2014.

[Lu *et al.*, 2015] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *ICCV*, pages 990–998, 2015.

[Luo *et al.*, 2011] Wei Luo, Xiaogang Wang, and Xiaoou Tang. Content-based photo quality assessment. In *ICCV*, pages 2206–2213. IEEE, 2011.

[Lv *et al.*, 2021] Pei Lv, Jianqi Fan, Xixi Nie, Weiming Dong, Xiaoheng Jiang, Bing Zhou, Mingliang Xu, and Changsheng Xu. User-guided personalized image aesthetic assessment based on deep reinforcement learning. *arXiv:2106.07488*, 2021.

[Ma *et al.*, 2017] Shuang Ma, Jing Liu, and Chang Wen Chen. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *CVPR*, pages 722–731, 2017.

[Mai *et al.*, 2016] Long Mai, Hailin Jin, and Feng Liu. Composition-preserving deep photo aesthetics assessment. In *CVPR*, pages 497–506, 2016.

[Murray *et al.*, 2012] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*. IEEE, 2012.

[O'Donovan *et al.*, 2011] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. Color compatibility from large datasets. In *SIGGRAPH*, pages 1–12. 2011.

[Ren *et al.*, 2017] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran. Personalized image aesthetics. In *ICCV*, pages 638–647, 2017.

[She *et al.*, 2021] Dongyu She, Yu-Kun Lai, Gaoxiong Yi, and Kun Xu. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In *CVPR*, pages 8475–8484, 2021.

[Sheng *et al.*, 2018] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. Attention-based multi-patch aggregation for image aesthetic assessment. In *ACMMM*, 2018.

[Talebi and Milanfar, 2018] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *TIP*, 27(8):3998–4011, 2018.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[Wang *et al.*, 2018] Guolong Wang, Junchi Yan, and Zheng Qin. Collaborative and attentive learning for personalized image aesthetic assessment. In *IJCAI*, 2018.

[Zeng *et al.*, 2019] Hui Zeng, Zisheng Cao, Lei Zhang, and Alan C Bovik. A unified probabilistic formulation of image aesthetic assessment. *TIP*, 29:1548–1561, 2019.

[Zhou *et al.*, 2017] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 40(6):1452–1464, 2017.

[Zhu *et al.*, 2020] Hancheng Zhu, Leida Li, Jinjian Wu, Sicheng Zhao, Guiguang Ding, and Guangming Shi. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *TCYB*, 2020.