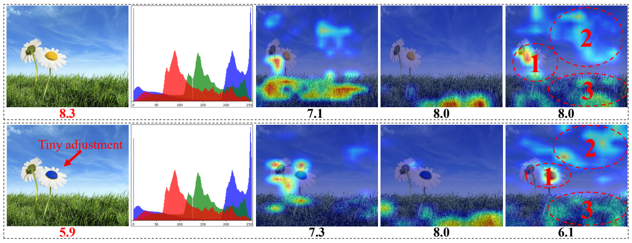
Delegate Transformer for Image Color Aesthetics Assessment

Anonymous ICCV submission

Paper ID ****

Abstract

We present a comprehensive study on a new task named image color aesthetics assessment (ICAA), which aims to assess color aesthetics based on human perception. ICAA is important for various applications such as imaging measurement and image analysis. However, due to the highly diverse aesthetic preferences and numerous color combinations, ICAA presents more challenges than conventional image quality assessment tasks. To advance ICAA research, 1) we propose a baseline model called the Delegate Transformer, which not only deploys deformable transformers to adaptively allocate interest points, but also learns human color space segmentation behavior by the dedicated module. 2) We elaborately build a color-oriented dataset, ICAA17K, containing 17K images, covering 30 popular color combinations, 80 devices and 50 scenes, with each image densely annotated by more than 1,500 people. 3) We develop a large-scale benchmark of 15 methods, the most comprehensive one thus far based on two datasets, SPAQ and ICAA17K. Our work, not only achieves state-of-the-art performance, but more importantly offers the community a roadmap to explore solutions for ICAA. Code and dataset are available in the supplementary material.



(a) Input (b) Color Histogram (c) MaxVit (d) NIMA (e) Ours

Figure 1. The results of different methods on two similar images with tiny color adjustment, ground-truth (red) and predicted (black) scores are shown below. (a) Similar images. (b) Color histogram [16] is unable to differentiate and quantify the color aesthetics. The saliency maps of (c) MaxViT [17], (d) NIMA [18] and (e) our method, whereas our method achieves a fine-grained perception of color aesthetics.

1. Introduction

Color is known as having a higher degree of discriminability and correlation compared to other visual features [1, 2, 3], because the human eye can directly perceive light of different wavelengths and convert them into different color signals. As digital photography expands rapidly, ICAA has become one of the most important criteria to automatically assess whether the image meets users' aesthetic preferences [4, 5, 6, 7]. It is also an essential step in imaging measurements among manufacturers to evaluate the performance of smartphones and cameras [8, 9, 10].

Although ICAA is an important branch of IAA (image aesthetics assessment), they deal with different tasks. IAA evaluates the holistic aesthetics of an image, which implicitly depends on color and other attributes (brightness,sharpness, etc.) [11, 12, 8]. ICAA focuses more on evaluating the impact of color harmony, color combinations and other factors on color perception.

ICAA is confronted with two major challenges. First, the types and combinations of colors are intrinsically complex, thus the color aesthetics are determined by specific colors and their relative positions in color spaces [13]. However, biological studies show that humans can perceive only up to 8 colors due to the human eye and perception limits [14]. Therefore, extracting dominant colors from an image is a crucial process and important prior knowledge [6, 15]. Second, human color preferences are subjective and vary from person to person based on factors such as age and culture that may change over time.

Most traditional quantization methods for ICAA rely on objective rules such as color histogram [16] or color wheel [19] theories. However, the color information obtained from pixel statistics is limited to capture the nuances of color aesthetics (Fig. 1 (b)), mainly due to these methods cannot perceive the spatial information of colors and the effect of image content and semantics. In addition, they are suited to qualitative analysis instead of directly quantifying the aesthetics of colors.

IAA methods, especially leaning-based IAA approaches such as [18, 20, 17], are not specifically designed for ICAA and then lack of prior color knowledge, which result in struggling to achieve a fine-grained perception of the importance of different color spaces, and being susceptible to interference from multiple visual elements (Fig. 1 (c)(d)). Furthermore, to our knowledge, there are no datasets specifically designed for ICAA, which results in most data-driven methods not generalizing well to deal with subjective ICAA tasks. In addition, the community of image quality assessment lacks of a fair ICAA benchmark for reliable comparisons.

The contributions of our paper lie in:

- The limits of human perception and pixel-based extraction methods for dominant colors are revealed, which guides us to propose a baseline model, called the Delegate (Deformable Gate) Transformer, to adaptively allocate interest points for dominant colors and simulate human behavior of color space segmentation.
- To verify our method more convincingly, a dataset, named ICAA17K, is exclusively designed for ICAA. Specifically, it is a ***color-oriented*** dataset with 17K images and the richest annotations thus far. Furthermore, dedicated preassessment strategy is adopted to alleviate long-tailed distributions of dataset.
- Based on ICAA17K, 15 state-of-the-art baselines (see Table 1) are evaluated, which takes our benchmark as the most complete one for ICAA thus far. Our work, not only gains state-of-the art performance, but more importantly serves as a potential catalyst for promoting large-scale model comparisons in future ICAA research.

2. Related Work

2.1. Image Color Aesthetics Assessment

Traditional quantization methods. To mitigate the gap between subjective and objective evaluations, early methods usually extracted hand-crafted features from brightness, hue, and saturation or applied color difference formulas. For instance, Yang *et al.* [11] employed the CIEDE2000 color difference formula to align objective evaluation results with subjective visual perception. Shi *et al.* [30] utilized hue as the primary color information and proposed the structure and hue similarity model (SHSIM) to assess image color quality. Some other works believed that aesthetically pleasing colors typically conform to certain established rules. O'Donovan *et al.* [6, 31] studied color compatibility theories and color aesthetics based on color pattern (templates). Cohen *et al.* [19] evaluated color harmony using a color wheel, while Nishiyama *et al.* [32] applied a color harmony model to evaluate the distributions of hue, lightness, and chroma.

However, the above-mentioned methods are based on statistical quantitative information of image pixels, ignoring how spatial and semantic content affect color aesthetics. Although these methods can give qualitative analysis results, they cannot quantify the aesthetic differences brought about by a tiny change in color, as shown in Fig. 1.

Data-driven methods. In the deep learning era, some data-driven methods are proposed to fit aesthetic information in images (Table 1). For example, the works of [18, 20, 28, 29] introduced additional context information, e.g., theme and layout information to improve the performance, while others developed models of personalized image aesthetics assessment (PIAA) [22, 33]. The latest works have explored adopting the transformer structure to extract more powerful aesthetic features [17, 27].

Nevertheless, these methods typically extract holistic aesthetic features and lack of prior color knowledge, which take themselves harder to perceive the spatial distribution and composition of different colors in an image, then leading to diffuse attention against perceiving color space. On the other hand, they cannot assign different attention weights based on color importance, which leads to in poor fine-grained perception for color (Fig. 5).

As revealed in Section I, although images are naturally composed of distinct color spaces, dominant and secondary colors have different effects on color perception. The proposed Delegate Transformer learns to segment color space from dedicated deformable attention rather than static pixel values, and thus captures spatial information of color. Furthermore, different color spaces are assigned different levels of attention by the Delegate Transformer, which exactly matches human behavior for color space segmentation.

2.2. Image Aesthetics Assessment Datasets

In recent years, several datasets have been developed for IAA, as shown in Table 2. Murray *et al.* [34] created the AVA dataset, a large-scale IAA dataset containing nearly 255,000 images that has become one of the most popular IAA datasets. However, this dataset does not provide annotations for color. Kong *et al.* [12] introduced the Aesthetics and Attributes Database (AADB), which includes subjective aesthetic scores and attributes annotated by individual users and provides two binary labels for color harmony. Yu *et al.* [35] developed the Photo Critique Captioning Dataset (PCCD) to address the problem of generating captions for photos based on their aesthetics, which includes evaluations of the color & lighting attributes. Recently, Fang *et al.* [8] presented the Smartphone Photography Attribute and Quality (SPAQ) dataset with colorfulness attribute scores.

The above datasets primarily focus on evaluating the holistic quality of images and lack of detailed color annotations, with limited color types or combinations. Specifically, these datasets exhibit serious selection bias. e.g., about 50% images of the AVA dataset are "black and white" images, which outnumber other colors by 10 to 100 times, and the PCCD and SPAQ datasets have few images of "pink" and "violet" colors. Therefore, these IAA datasets

No.	Model	Year	Pub.	Basic	Platform
1	RAPID [21]	2014	ACMMM	incorporate heterogeneous	Python&Lua
2	AADB [12]	2016	ECCV	sampling strategy, ranking loss	MATLAB
3	PAM [22]	2017	ICCV	residual-based, active learning	Caffe
4	A-Lamp [23]	2017	CVPR	layout-aware, multi-patch	Scipy
5	NIMA [18]	2017	TIP	predict distribution	Tensorflow
6	MP$_{ada}$ [24]	2018	ACMMM	attention, multi-patch	Tensorflow
7	MLSP [20]	2019	CVPR	staged training,multi-level features	Tensorflow
8	MT-A [8]	2020	CVPR	multi-task, high-level semantics interact	PyTorch
9	BIAA [25]	2020	TCYB	meta-learning, bilevel optimization	PyTorch
10	UIAA [26]	2020	TIP	unified probabilistic formulation	MATLAB
11	MUSIQ [27]	2021	ICCV	multi-scale representation	PyTorch
12	HGCN [28]	2021	CVPR	graph convolution networks	Jittor
13	TANet [29]	2022	IJCAI	fuse theme adaptively	PyTorch
14	MaxViT [17]	2022	ECCV	multi-axis attention	PyTorch
15	Ours	2023	ICCV	delegate attention, color space segmentation	PyTorch

Table 1. Summary of 15 existing representative IAA or PIAA models and the proposed methods. **Code**: available code links are marked in red, and our code is provided in the supplementary material.

Dataset	Year	Pub.	Images	Rating
DP Challenge [36]	2008	ICIP	16,509	100
CUHK-PQ [37]	2011	ICCV	17,673	10
AVA [34]	2012	CVPR	**255,530**	250
AADB [12]	2016	ECCV	10,000	5
PCCD [35]	2017	ICCV	4,235	7
DPC-Captions [38]	2019	ACMMM	154,384	6
SPAQ [8]	2020	CVPR	11,125	600
TAD66K [29]	2022	IJCAI	66,327	1,200
ICAA17K (Ours)	2023	ICCV	17,726	**1,500**

Table 2. Summary of the related datasets.

are not suitable for ICAA tasks and cannot support the generalization of ICAA models well.

To address the above issue, we try to develop a specialized and color-oriented dataset the first time. To the best of our knowledge, our ICAA17K dataset is the largest, as well as most densely annotated ICAA dataset with a diverse range of color types and image collection devices (see Table 2). It also has a dedicated dataset construction strategy to avoid selection bias and long-tailed distributions, which ensures it a high-quality and robust dataset.

3. Proposed Approach

Dominant or primary colors tend to draw attention quickly and are often the focal point of visual design. Instead of using traditional methods that rely on the number of pixels to differentiate between dominant and secondary colors, we directly employ attention mechanism to distinguish them in a visual context. First, a sparse, data-related, and deformable attention mechanism are introduced in our baseline model to allocate the attention by setting sparse interest points. Second, the color space segmentation module is utilized to group these interest points into different color spaces, and determine the importance of each space.

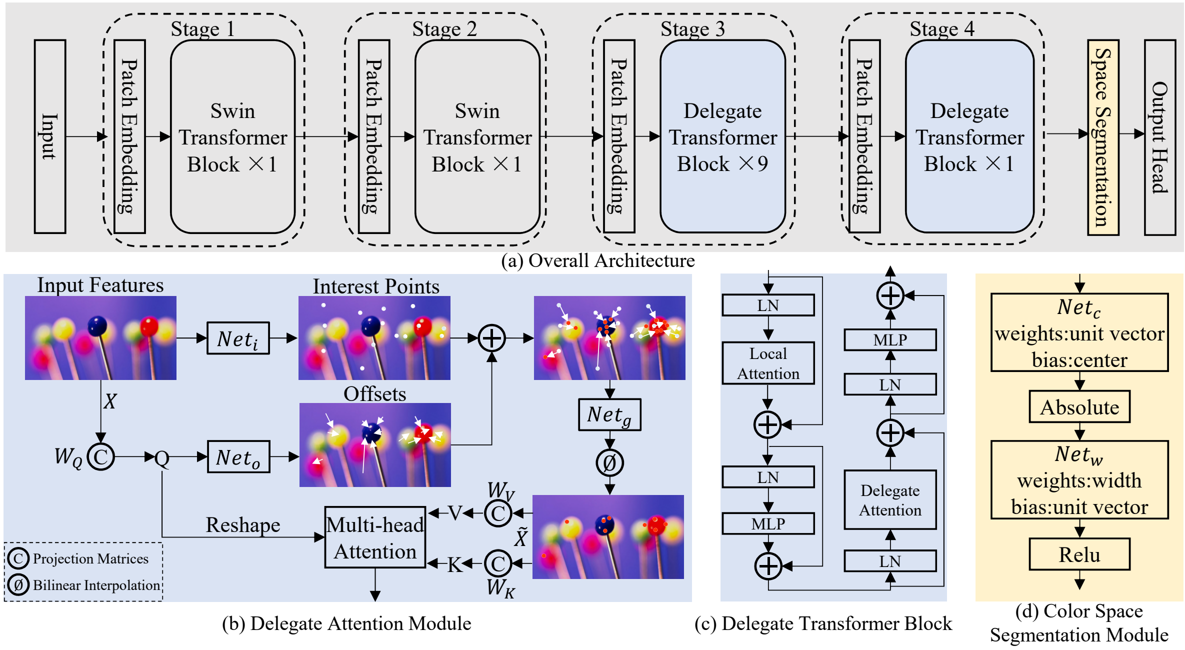
Overall architecture. An overview of the architecture is presented in Fig. 2 (a). Our implementation begins with embedding an input image through a 4×4 convolution operation. The first and second stages consist of two Swin Transformer blocks [39]. At these stages, the feature maps are processed by shifting window attention to locally aggregate aesthetic information. Subsequently, several proposed Delegate Transformer blocks, which form Stage III and IV, are utilized to process these feature maps globally and model the long-range relationships. Finally, these feature points are sent to a dedicated module to divide the color space and then sent to an output head to predict the color aesthetics score.

3.1. Delegate Transformer for Allocating Attention

Utilizing the projection matrices to generate queries Q, keys K, and values V from the original feature map X, the classical multihead self-attention (MSA) can be calculated as:

$$\mathrm{MSA} = \mathrm{Softmax}\left(\mathbf{Q}^{(m)}\mathbf{K}^{(m)T}/\sqrt{D}\right)\mathbf{V}^{(m)T}, \qquad (1)$$

where D is the query/key dimension, m denotes the m-th attention head. However, the calculation based on the entire feature map significantly increases the computational



(a) Overall Architecture

(b) Delegate Attention Module

(c) Delegate Transformer Block

(d) Color Space Segmentation Module

Figure 2. The proposed model. (a) Overall architecture. (b) Feature processing flow. The input features are sent to learn the original interest points and adjusted by offsets, then filtered by the gate module. (d) The entire block. (d) The color space segmentation module, which is modeled as a stack of CNN layers.

complexity. The output attention map also requires a large number of calculations when sent to the next block for color space segmentation. Instead of the calculation for the entire feature map, in this paper, the corresponding calculation is delegated to interest points for achieving fewer computations and higher flexibility (Fig. 2 (b)).

The feature map X is fed into a lightweight subnetwork Net_i to generate the initial interest points. Then, another subnetwork Net_o is adopted to calculate the offset from Q to adjust the positions of the interest points, which is formulated as:

$$\begin{aligned} &Q = XW_Q, K = \widetilde{X}W_k, V = \widetilde{X}W_v, with \\ &X_i = \O[X; G * (\underbrace{Net_i(X)}_{\text{interest points}} + \underbrace{Net_o(Q)}_{\text{offsets}})], \end{aligned} \quad (2)$$

where interest points and offsets are both represented by two-dimensional coordinates that are normalized to the range [-1, +1], where (-1, -1) represents the top-left corner and (+1, +1) represents the bottom-right corner. A gating module, G, is employed to determine activation of the current adjusted interest point coordinates, reducing redundancy further in the space and preventing outliers and overlapping interest points. It can be expressed as:

$$G = Sigmoid(Net_g(X) \cdot \alpha), \quad (3)$$

where α is a value large enough to ensure that the gate matrix is either 0 or 1. After obtaining the adjusted and filtered interest points $\widetilde{X}$, the bilinear interpolation is applied to sample the output feature from $\widetilde{X}$, which can be represented as

$$\O[X;(p_x,p_y)] = \sum_{(q_x,q_y)} M(p_x,p_y)M(q_x,q_y)X(q_x,q_y), \quad (4)$$

where (p_x, p_y) represents one of the coordinates of the interest points, $M(a,b) = max(0, 1 - |a - b|)$, and (q_x, q_y) is non-zero on only the 4 integral points in X closest to (p_x, p_y). Finally, the standard multihead attention mechanism of formula (1) is applied to calculate the output.

3.2. Module for Color Space Segmentation

The selection of dominant and secondary colors is based on human understanding of the relationship between image content and color, which is highly subjective and leads to uncertainty in the size and segmentation of color spaces. To address this issue, we propose a learning-based approach to determine the size and the segmentation of color spaces. After obtaining the color features of the interest points, 1) the interest points are grouped into several color spaces (κ pre-defined types). 2) Each color space is mapped to the [0, +1] interval, where each interval can be determined by its widths and centers. 3) By counting the weight distribution of points in each color space, it is easy to distinguish the attention of each color space, thus dividing the dominant and secondary color space.

However, using fixed bin widths and centers is not differentiable, and the parameters cannot be updated through

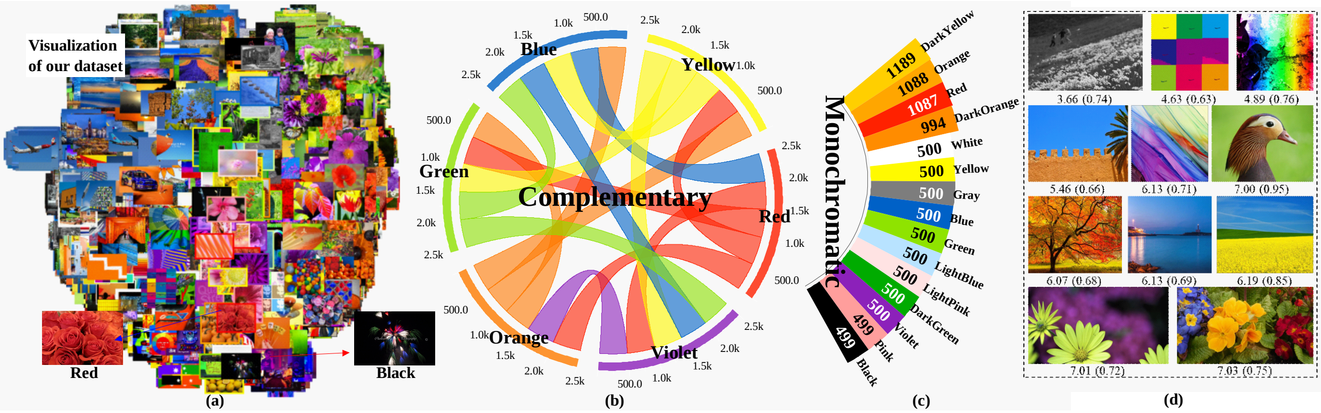


Figure 3. The proposed ICAA17K dataset. (a) Visualization of images with different color types; (b) quantification of 15 complementary/polychromatic colors; (c) quantification of 15 monochromatic colors; (d) examples in the proposed ICAA17K dataset with the ground truth (and confidence) scores shown below each image.

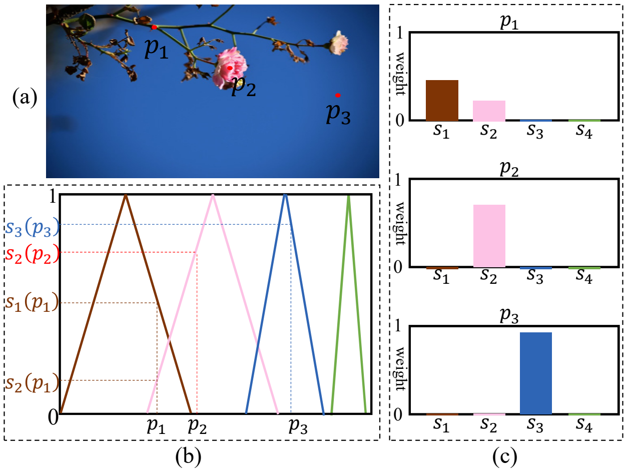


Figure 4. An example of color space segmentation with 4 color spaces ($\kappa = 4$). (a) Three sample points. (b) The weights of the three points in each space calculated through formula (5). (c) The weights of each point in each space.

backpropagation. To resolve this issue, the widths and centers are adaptively generated using two lightweight networks, Net_w and Net_c, which can be learned during training. To calculate the uncertainty in color space segmentation, our model learns a probability distribution for each point p_j to indicate its weight in κ color spaces, denoted as $S(p_j) = [s_1(p_j), ..., s_\kappa(p_j)]$, the weight $s_i()$ of p_j allocated to the i-th space is determined by the piecewise linear basis function.

$$s_i(p_j) = \max\left\{0, 1 - w_i \times |p_j - c_i|\right\}, \qquad (5)$$

where w_i and c_i are the widths and centers of the i-th color space. If an interest point falls into the i-th bin, e.g., the interval of $[c_i - w_i, c_i + w_i]$, it votes for the i-th bin with a positive weight $s_i(p_j)$. Fig. 4 shows an example with $\kappa = 4$ color spaces.

The operation in formula (5) can be designed into a lightweight color space partition block (Fig. 2 (d)), where $p_j - c_i$ is equivalent to convolving the point feature maps with a fixed 1×1 kernel and a learnable bias c_i. Then, after taking the absolute value, the remaining operation is completed using another learnable weight, a fixed 1×1 bias, and a ReLU activation function.

4. Proposed Dataset

4.1. Image Collection

How to avoid selection bias by building a comprehensive ICAA dataset. Selection bias occurs when some types of the intended color, scenes and devices are less likely to be included than others, it can affect the validity and generalizability of training models. To solve this issue, 1) we analyzed the most frequently uploaded colors on the Flickr and COLOURLovers website from 2009 to 2022. Finally, 15 complementary/polychromatic colors and 15 monochromatic colors are selected as the basic color types, which ensures that the dataset includes a diverse range of color types (Fig. 3). 2) To improve the dataset's generalizability across different scenes, the images in ICAA17K cover more than 50 scene categories, e.g., animals, plants, mountains, landscapes, and night scenes. 3) We collected images taken by more than 80 different types of devices, including mobile phones (e.g., iPhone X, Huawei Mate 10, and Samsung Galaxy S10) and digital cameras (e.g., Canon, Nikon, Sony, Panasonic), to encompass a wide range of devices.

How to alleviate long-tailed distributions by performing a preassessment strategy. The annotated label in the most aesthetic datasets have long-tailed distributions [34, 8], which bias models trained on imbalanced data toward majority examples. To avoid correcting the long-tailed distributions after label annotation, we propose a preassess-

Metric [ICAA]	2014-2019							2020-2022								
	RAPID [21]	AADB [12]	PAM [22]	A-Lamp [23]	NIMA [18]	MP$_{ada}$ [24]	MLSP [20]	MT-A [8]	BIAA [25]	UIAA [26]	MUSIQ [27]	HGCN [28]	TANet [29]	MaxViT [17]	Ours *1	Ours *2
SRCC↑	.759	.788	.796	.810	.809	.810	.826	.821	.820	.817	.835	.826	.829	.855	.873	**.887**
LCC↑	.771	.793	.804	.818	.815	.819	.837	.830	.829	.830	.841	.831	.836	.874	.890	**.901**
Accu↑	.836	.887	.899	.908	.906	.906	.921	.908	.913	.910	.918	.909	.903	.920	.957	**.965**
Epoch↓	80	60	60	50	60	50	50	60	100	60	80	60	110	200	**25**	40

Table 3. Comparisons of 15 models on our ICAA17K dataset. We ***retrained the models for the best performance*** with the recommended parameter and dataset settings. See subsection 5.1 for details regarding the metrics. “Ours *1” indicates that, similar to other methods, only the color annotation scores from the ICAA17K dataset are used during training. “Ours *2” indicates that the holistic aesthetics score is jointly predicted when using multiple tasks.

Metric [SPAQ]	2014-2019							2020-2022							
	RAPID [21]	AADB [12]	PAM [22]	A-Lamp [23]	NIMA [18]	MP$_{ada}$ [24]	MLSP [20]	MT-A [8]	BIAA [25]	UIAA [26]	MUSIQ [27]	HGCN [28]	TANet [29]	MaxViT [17]	Ours * 1
SRCC↑	.721	.725	.732	.747	.730	.746	.753	.751	.748	.750	.753	.755	.750	.753	.770
LCC↑	.730	.739	.744	.760	.746	.757	.773	.763	.759	.763	.770	.771	.761	.770	.791
Epoch↓	95	75	70	65	55	65	55	30	85	70	90	45	110	150	25

Table 4. Comparisons of 15 models on the SPAQ dataset [8]. All models used only the colorfulness score in the dataset for training, and ***were retrained for the best performance*** using the recommended parameters and dataset settings.

ment strategy during data collection to maximize label diversity.

First, the color harmony [19] is applied to evaluate objective color quality by measuring the distance between the tested image and templates on the hue wheel. Given a harmonic template $T(\beta)$ (one of the predefined type of templates T) with an associated orientation β, the color harmony f_{obj} of a tested image C could be calculated as follows:

$$f_{obj}(C) = \sum_{l \in C} \left\| H(l) - E_{T(\beta)}(l) \right\| \cdot Y(l), \tag{6}$$

where l is the hue of a pixel, $E_{T(\beta)}$ is the sector border hue of harmonic template $T(\beta)$, H and Y mean the hue and the saturation channels, respectively, $\|\cdot\|$ denotes the arc-length distance on the hue wheel (measured in radians).

Second, we crawl images from Flickr and record their metadata, including “views” (number of visits VI) and “faves” (number of clicks that favor image FA). Intuitively, an image with a higher ratio of faves to views is more popular and beautiful. Specifically, considering the affect of user’s herd mentality, we adopt the following formula to approximate the subjective perception:

$$f_{subj}(C) = \frac{FA(C)}{\sqrt{VI(C)}}. \tag{7}$$

Finally, we normalize the f_{subj} and f_{obj} , and add them together as the final preassessment score of the image (not shown to annotators), and extract balanced samples from each score range for the next stage of manual evaluation.

4.2. Image Annotation

How to avoid unreliably by improving annotation quality. The traditional annotation process suffers from a lack of a baseline or reference for evaluating aesthetics, which makes subjective responses unreliable in the early stages. To mitigate this problem, 1) we first provided annotators simple guidelines based on photography theories to guide their judgment of image color aesthetics. 2) We provided anchor images with different rating scales as references. Experts rated the anchor images based on their rich photography experience, allowing annotators to understand the differences between rating scales. Before annotating, annotators had to review the anchor images and guidelines thoroughly. The average value was then calculated as the ground truth. 3) Annotators could provide a confidence score (from 0 to 1) to indicate their certainty in each rating. This information can be useful for studying the noise and subjectivity in the dataset. 4) Finally, each image has an annotation score from 1 to 10, which represents color aesthetic perception from lowest to highest, we selected images with an average confidence score above 0.5 as samples in the dataset. We collected about 1,500 opinions for each image. Some examples are shown in Fig. 3 (d).

5. Experiments

5.1. Experimental Settings

Benchmark Datasets. We performed model evaluations on two datasets, the proposed ICAA17K dataset and the SPAQ

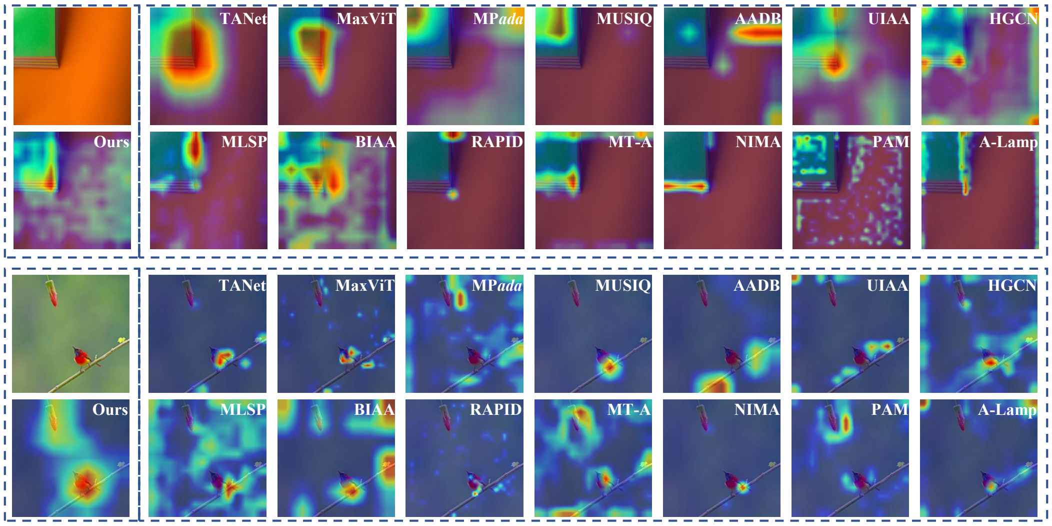


Figure 5. Saliency maps of the 15 models. The attention of our method is segmented into distinct regions based on the color distribution.

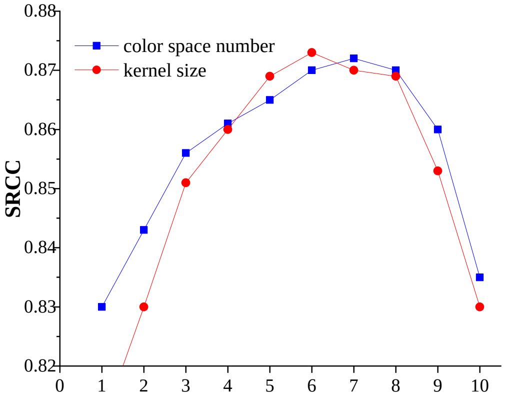


Figure 6. Analysis of kernel sizes in Net_i and the number of color spaces κ. Larger kernel sizes result in fewer interest points.

dataset [8]. Although the SPAQ dataset focuses on image quality rather than aesthetics, it provides an annotation on colorfulness, which can also afford testing the ability of various baseline models to extract color features to some extent.

Benchmark Models and Training Protocols. To the best of our knowledge, there is no publicly available learning-based ICAA model. Therefore, we selected 14 deep learning baselines (Table. 1) according to the following criteria: 1) classical architectures with available code and 2) SOTA in a specific field, e.g., personalized IAA. These baselines are trained with the recommended parameter settings (e.g., optimizer and batch size), using the same training and testing settings. The mean squared error (MSE) is adopted as the loss function across all methods.

Evaluation Metrics. We adopt two popular evaluation metrics, the Spearman's rank correlation coefficient (SRCC) and the linear correlation coefficient (LCC), to evaluate the model's ability to perform fine-grained evaluations. The metrics for the ICAA17K dataset also include binary classification accuracy (color aesthetically negative or positive). Since the ground-truth annotations in the ICAA17K dataset are based on a 5-point boundary line, if an annotator tends to give a score greater than 5, it is considered aesthetically positive; otherwise, it is considered negative. Therefore, this metric can evaluate the model's ability to perform coarse-grained evaluations.

5.2. Results and Dataset Analysis

ICAA17K Dataset. Table 3 lists the results of 15 models on the ICAA dataset. The proposed baseline method (marked by "Ours *1") achieves the best performance on all metrics and has a higher training speed. This is mainly due to its sparse attention and adaptive color space partitioning. Furthermore, our ICAA17K dataset also provides additional annotations regarding holistic aesthetics. Therefore, we incorporate this additional supervision information into the multitask training (marked by "Ours *2"), which improves performance but increases the training time.

SPAQ Dataset. Based on the performance of the 15 methods reported in Table 4, our method still achieves the top performance on all metrics with fewer training epochs, showing that it is a promising solution for the IQA problem.

Dataset Analysis. Compared to the SPAQ dataset, when using the ICAA17K dataset for training, most networks converge faster and have higher SRCC and LCC metrics, which indicates that **models trained on the ICAA17K dataset more accurately fit human subjective color perception**. It also suggests that introducing more annotators (SPAQ 600 *vs* ICAA17K 1500) and providing more annotation references for annotators can alleviate noise caused by subjective

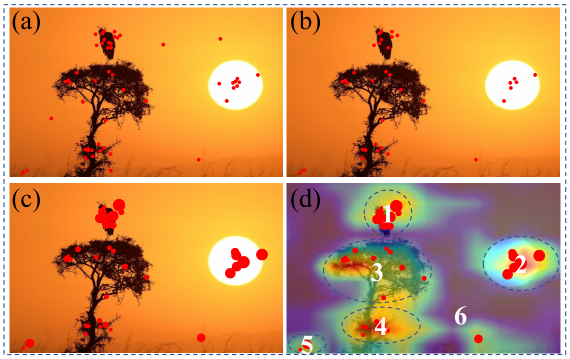


Figure 7. Visualization of our attention mechanism. (a) Sampled interest points adjusted by offsets. (b) Interest points filtered by the gating module. (c) The radius of the circles represents the attention weights at multiple heads, with a larger radius indicating a higher weight. (d) The color space segmentation of each point, where the numbers represent the primary and secondary relationship of the color space, e.g., "1" represents the dominant color.

annotation, which is a promising direction toward improving the quality of IQA or IAA datasets.

5.3. Ablation Study

Delegate *vs* Shift/Deformable Transformers. In Table 5, we replace the Delegate Transformer blocks with the shift attention mechanism of Swin Transformer [39] and deformable attention of DAT Transformer [40] in the last stage. In this case, **our model has better performance and lower computation**, which outperforms Swin by 5.2% in the SRCC with 41.7% FLOPs and outperforms DAT by 4.4% in the SRCC with 41.7% FLOPs. Additionally, the stacking way also has a significant impact on the performance. The Delegate Transformer blocks may result in information loss if it is employed in the early stages due to the sparse attention mechanism.

The Number of Interest Points and Color Space. The number of interest points mainly depends on the kernel size of the subnetwork Net_i. In Fig. 6, the results show that **an inadequate or excessive number of interest points leads to degradation in the model's performance**. Moreover, **selecting the color space hyperparameter κ is also crucial for the performance** of our model. Our experiments reveal that the optimal performance is achieved when $\kappa = 6$. Additionally, a larger value of κ leads to an increase in computation and memory requirements. In contrast, a smaller value of κ may not provide a sufficient characterization of the color composition in the input image.

Different Modules. The effectiveness of the offset generation module, gating module, and color space segmentation module is evaluated in Table 6. It is observed that without the contribution of these modules, the SRCC of the baseline model is reduced by 2.5%, 1.5%, and 3.4%, respectively. It is worth mentioning that **these modules incur only slight increases in computational cost**.

	Stages				FLOPs↓	SRCC↑	LCC↑
#	1	2	3	4			
1	S	S	S	S	240G	0.830	0.843
2	S	S	D	D	240G	0.836	0.848
3	S	S	G	G	140G	**0.873**	**0.890**
4	G	G	S	S	210G	0.828	0.835
5	S	G	G	S	140G	0.838	0.850
6	G	S	S	G	210G	0.825	0.833
7	G	G	G	G	**130G**	0.837	0.846

Table 5. Ablations on the application of different attention mechanisms in different stages, where *S*, *D* and *G* represent the application of Swin, DAT and our blocks, respectively.

Method	FLOPs ↓	SRCC↑	LCC↑
baseline	140G	**0.873**	**0.890**
no Net_o	**139.97G**	0.851	0.875
no Net_g	**139.97G**	0.860	0.877
no space seg.	139.993G	0.843	0.855

Table 6. Ablations on the performance of different modules.

5.4. Visualization

The GradCAM [41] method is applied to visualize the saliency maps in Fig. 5. Our model differs from other methods because it segments the image into distinct regions based on the color distribution, and the region that typically receives the greatest attention is the region where the dominant color is located. The reason for this phenomenon is further explained in Fig. 7. First, our model generates interest points via Net_i and adjusts their positions through Net_o. Second, some redundant points are eliminated via Net_g and weights are calculated using multihead attention. Third, the color space segmentation module and formula (5) are employed to divide the different color spaces and direct different attention toward them.

6. Conclusions

In this paper, a comprehensive study on ICAA has been conducted by developing a Delegate Transformer model as a baseline, constructing a color-oriented dataset benchmark, and establishing the largest-scale benchmarks to date. Compared with traditional or data-driven models, our model achieves superior performance and produces visually consistent results. We hope our work should contribute to the community's research on ICAA. However, ICAA remains far from being solved, we will continue to optimize our approach by incorporating more color prior knowledge in the future.

References

[1] Ahmed Talib, Massudi Mahmuddin, Husniza Husni, and Loay E George. A weighted dominant color descriptor for content-based image retrieval. *Journal of Visual Communication and Image Representation*, 24(3):345–360, 2013.

[2] Yining Deng, BS Manjunath, Charles Kenney, Michael S Moore, and Hyundoo Shin. An efficient color representation for image retrieval. *IEEE Transactions on image processing*, 10(1):140–147, 2001.

[3] Bangalore S Manjunath, J-R Ohm, Vinod V Vasudevan, and Akio Yamada. Color and texture descriptors. *IEEE Transactions on circuits and systems for video technology*, 11(6):703–715, 2001.

[4] Luca Marchesotti, Naila Murray, and Florent Perronnin. Discovering beautiful attributes for aesthetic image analysis. *International journal of computer vision*, 113:246–266, 2015.

[5] Andrew J Elliot and Markus A Maier. Color psychology: Effects of perceiving color on psychological functioning in humans. *Annual review of psychology*, 65:95–120, 2014.

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

[7] Joon-Young Lee, Kalyan Sunkavalli, Zhe Lin, Xiaohui Shen, and In So Kweon. Automatic content-aware color and tone stylization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2470–2478, 2016.

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

[9] Emilie Baudin, François-Xavier Bucher, Laurent Chanas, and Frédéric Guichard. Dxomark objective video quality measurements. *Electronic Imaging*, 2020(9):166–1, 2020.

[10] Sabine E Susstrunk and Stefan Winkler. Color image quality on the internet. In *Internet imaging V*, volume 5304, pages 118–131. SPIE, 2003.

[11] Yang Yang, Jun Ming, and Nenghai Yu. Color image quality assessment based on ciede2000. *Advances in Multimedia*, 2012, 2012.

[12] 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.

[13] Xuan Xu and Xin Li. Scan: Spatial color attention networks for real single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[14] Aleksandra Mojsilovic, H Hu, and Emina Soljanin. Extraction of perceptually important colors and similarity measurement for image matching, retrieval and analysis. *IEEE Transactions on Image Processing*, 11(11):1238–1248, 2002.

[15] Stephen E Palmer, Karen B Schloss, and Jonathan Sammartino. Visual aesthetics and human preference. *Annual review of psychology*, 64:77–107, 2013.

[16] Michael J Swain and Dana H Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991.

[17] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *ECCV*, 2022.

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

[19] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. Color harmonization. In *ACM SIGGRAPH 2006 Papers*, pages 624–630. 2006.

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

[21] 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.

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

[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 *CVPR*, pages 722–731, 2017.

[24] 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.

[25] 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.

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

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

[28] 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.

[29] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. *IJCAI*, 2022.

[30] Yunyu Shi, Youdong Ding, Ranran Zhang, and Jun Li. Structure and hue similarity for color image quality assessment. In *2009 International Conference on Electronic Computer Technology*, pages 329–333. IEEE, 2009.

[31] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. Collaborative filtering of color aesthetics. In *Proceedings of the Workshop on Computational Aesthetics*, pages 33–40, 2014.

[32] Masashi Nishiyama, Takahiro Okabe, Imari Sato, and Yoichi Sato. Aesthetic quality classification of photographs based on color harmony. In *CVPR 2011*, pages 33–40. IEEE, 2011.

[33] 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.

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

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

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

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

[38] 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.

[39] 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.

[40] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4794–4803, 2022.

[41] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.