Thinking Aesthetics Assessment of Image Color Temperature: Models, Datasets and Benchmarks

Jinguang Cheng[1*], Chunxiao Li[2*], Shuai He[1†], Taiyu Chen[1], Anlong Ming[1†]

[1]School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
[2]School of Artificial Intelligence, Henan University, Zhengzhou, China
jgcheng@bupt.edu.cn, chunxiaol@henu.edu.cn, {hs19951021, 2457319307, mal}@bupt.edu.cn

Abstract

Color temperature, as a crucial attribute influencing image color, plays a critical role in Image Aesthetics Assessment (IAA). Yet, within the existing IAA field, little light has been shed on assessing the aesthetic quality of image color temperature. To bridge this gap, we introduce a new task: Image Color Temperature Aesthetics Assessment (ICTAA). However, this task poses the following challenges: 1) **Perceptual Sensitivity**: humans exhibit high sensitivity to subtle shifts in color temperature, necessitating a model to enable fine-grained discrimination; 2) **Spectral Continuity**: The theoretical modeling of color temperature aesthetics requires continuous labels; however, the just-noticeable-difference property of human perception makes continuous labeling infeasible, necessitating a well-designed labeling strategy. To address the aforementioned challenges, we make the following efforts. First, we propose a multi-modal contrastive learning framework, ICTA2Net, that models color temperature differences between image pairs while strictly controlling other visual attributes. Second, leveraging color temperature transitivity, we design a weakly supervised strategy that discretely samples images based on anchor images and human perception to build contrastive relations across color temperatures, enabling learning from discrete labels. Thirdly, we construct a color temperature aesthetics dataset, ICTAA240K, and a benchmark for validation. Additionally, we propose a new metric, Information Entropy-weighted Accuracy (IEA), which weights accuracy by the degree of annotation disagreement to reflect model performance across varying sample difficulties, complementing existing evaluation metrics. Experiments show our method outperforms existing state-of-the-art IAA methods on ICTAA240K, thereby setting an effective roadmap for ICTAA.

Code — https://github.com/chasecjg/ICTA2Net

1 Introduction

With the proliferation of social media and the rapid advancement of intelligent imaging devices, images have become a critical medium for information transmission. The attention of users to images has gradually extended from basic image quality to include higher-level aesthetic quality. Among many attributes that affect image aesthetics, color temperature, as one of the most direct and fundamental attributes influencing color perception, has been widely shown to be closely related to human aesthetic perception (Valdez and Mehrabian 1994; Ou et al. 2004; Li et al. 2021; Jiang et al. 2022), playing a significant role in the aesthetic performance of images (Leder et al. 2022; DXOMARK 2024).

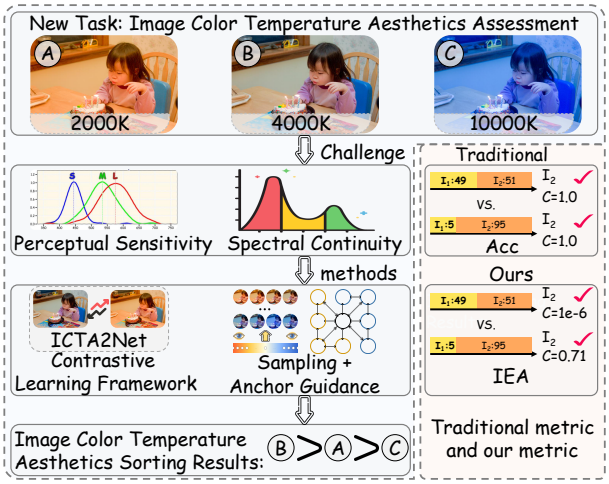

Figure 1: Illustration of the proposed ICTAA task. With respect to the overarching challenges of this task, perceptual sensitivity, and spectral continuity, we propose a contrastive learning framework that is guided by a weakly supervised strategy based on anchor images and enables the ranking of image color temperature aesthetics. The performance of the model is subsequently evaluated using the IEA metric.

In recent years, rapid advances in computer vision have driven substantial progress in the field of computational Image Aesthetics Assessment (IAA), spawning a series of excellent algorithms, such as NIMA (Talebi and Milanfar 2018), TANet (He et al. 2022), and BAID (Yi et al. 2023). Although these methods have achieved promising results, they remain inadequate for assessing images with varying color temperatures. This limitation stems primarily from the following challenges:

- **Perceptual Sensitivity**: Human vision exhibits sensitivity to shifts in color temperature, where even subtle shifts can lead to noticeable aesthetic perception changes. This

*These authors contributed equally.
†Corresponding author

necessitates models capable of fine-grained discrimination of color temperature differences (Liu et al. 2024). However, most of the IAA methods perform aesthetic assessments based on general attributes, failing to eliminate perception interference from other attributes (such as composition, exposure, and contrast) on image color temperature, thus making it difficult to focus solely on perceiving changes in image color temperature.

- **Spectral Continuity**: Color temperature, as a continuous physical quantity, theoretically requires modeling subtle aesthetic variations based on continuous supervisory labels. However, the Just-Noticeable-Difference (JND) nature of human perception makes continuous labeling infeasible (MacAdam 1942; Jiang et al. 2024). Existing IAA methods rely on heterogeneous images with specific attribute values to learn aesthetic representations, making it difficult for models to capture how aesthetic qualities vary with other visual attributes under controlled content, especially for continuously varying attributes like color temperature, which is also another vital cause for the model to remain insensitive to image color temperature variation.

To address the above challenges, we introduce a new task, Image Color Temperature Aesthetics Assessment (ICTAA), which aims to quantify the aesthetic performance of an image under varying color temperature shifts, grounded in human perception as shown in Fig. 1. Specifically, ICTA2Net is elaborately designed with the following aspects.

For the first issue, to finely perceive changes in image color temperature, we propose to use contrastive learning to focus only on the changes in color temperature by strictly controlling other visual attributes. And we simulate the human cognitive process in ICTAA, namely, first perceiving color temperature, and then making a multi-dimensional and comprehensive aesthetic comparison based on the image color temperature and content, constructed a modular aesthetic assessment framework for image color temperature based on multi-modal contrastive learning, named ICTA2Net. Furthermore, to enhance cross-modal semantic alignment, we also design an inter-group loss. **For the second issue**, our thinking is grounded in two key observations. First, although the light spectrum is continuous, human perception of color change operates under the principle of JND, that is to say, which allows for reasonable discretization at the perceptual level. Second, we observed that when the color temperature of an image changes in a consistent direction, its aesthetic perception exhibits a certain degree of **transitivity**, and this transitive characteristic was validated through a subjective experiment. Based on the two observations, we introduce an anchor image as a constraint reference and perform discrete sampling of images with color temperature shifts in the same direction, guided by human visual perceptibility. Then, we construct image pairs with an anchor image as a constraint for the sampled images. Based on these, we design a structured weakly supervised strategy.

To train our model, we construct a large-scale color temperature aesthetics dataset for the ICTAA task, containing over 240K images and corresponding text information. **Additionally**, we also propose a new metric, termed Information Entropy-weighted Accuracy (IEA), which weights accuracy based on the degree of annotation disagreement (derived from label distribution) to reflect model performance across samples of varying difficulty. The contributions of this paper are summarized as follows:

- To the best of our knowledge, this paper introduces the ICTAA task for the first time and discloses its key challenges: **Perceptual Sensitivity** and **Spectral Continuity**.
- To address the challenges, we propose a pairwise-based multi-modal contrastive learning framework, ICTA2Net, which models image color temperature aesthetics via image pairs contrastive learning by strictly controlling other attributes. In addition, we design a structured weakly supervised strategy rooted in anchor images and human visual perceptibility, which leverages intrinsic structural priors to construct supervisory sample pairs. We also construct a large-scale multimodal dataset, ICTAA240K, containing over 240,000 images under varying color temperature, with high-quality text descriptions. Additionally, we also introduce a new evaluation metric, IEA, to reflect model performance across samples of varying difficulty, complementing existing evaluation metrics.
- Based on the ICTAA240K dataset, we benchmark 10 representative methods, demonstrating that our proposed method achieves the state-of-the-art (SOTA) and provides a strong foundation for future research in ICTAA.

2 Related Works

As discussed in the previous section, the ICTAA task is novel, and to our knowledge, there is no directly related prior work. Nevertheless, two research areas are closely related to our work: White Balance and Related Evaluation and IAA.

2.1 White Balance and Related Evaluation

White balance plays a critical role in image color correction, aiming to eliminate color casts caused by variations in the color temperature of light sources, thereby restoring images to their visual appearance under standard white light illumination (Liu, Chan, and Chen 1995). Early methods often relied on simplified statistical assumptions such as the Gray World and Perfect Reflector hypotheses (Lam 2005), which struggle to handle the complexity and diversity of real-world scenes. With the advancement of deep learning, an increasing number of methods leverage neural networks to learn color correction parameters directly from images (Barron 2015; Bianco and Cusano 2019; Li et al. 2023; Li, Kang, and Ming 2023). Recently, some studies have attempted to incorporate image style factors to assist with white balance adjustment or to use user preferences as learning objectives (Kınlı et al. 2023; Afifi et al. 2025).

However, despite significant progress in physical accuracy achieved by these learning-based methods, their evaluation still primarily relies on objective physical metrics such as color difference $\Delta E2000$ and angular error (Sharma, Wu, and Dalal 2005), etc. These focus mainly on the numerical proximity between corrected images and "reference ground

truth", namely, color accuracy. Therefore, they do not fully reflect the fact that human preferences for color temperature are subjective and vary across different scenarios. As mentioned above, recently, some studies have considered user preferences as learning objectives (Afifi et al. 2025). Yet, few studies specifically model or explore the subjective aesthetic impact of color temperature variations, resulting in a disconnect between the physical objectives of color correction and the aesthetic quality of visual perception.

2.2 Image Aesthetics Assessment

IAA aims to quantify human subjective perception of the aesthetic quality of images. According to the assessment approach, existing research can be broadly categorized into two types: holistic IAA and multi-attribute IAA (Kong et al. 2016; Jin et al. 2019; Xie et al. 2024).

Holistic IAA focuses on predicting a holistic score through the holistic aesthetic quality of images. For example, Talebi *et al.* proposed the NIMA model, which predicts aesthetic score distributions to model subjectivity in aesthetic perception and remains a widely used benchmark (Talebi and Milanfar 2018). Recently, some methods attempt to improve holistic IAA via attention-based patch aggregation (Sheng et al. 2018), resolution-adaptive architectures (Hosu, Goldlucke, and Saupe 2019), and multimodal vision-language alignment (Wu et al. 2024), etc. However, this holistic IAA modeling paradigm suffers from limited interpretability, making it difficult to understand the influence of specific attributes on image aesthetics. Consequently, some researchers have begun exploring multi-attribute aesthetic assessment approaches.

Multi-attribute IAA focuses on the influence of multiple aesthetic attributes on the aesthetic quality of images, such as composition, color, contrast, clarity, and style. These methods are designed to model and assess the contribution of different visual attributes to image aesthetic quality. For example, Kong *et al.* constructed a large-scale dataset, AADB, which includes multiple aesthetic attribute annotations (e.g., rule of thirds, color harmony, etc.), and introduced a pairwise ranking loss to learn relative aesthetic preferences (Kong et al. 2016). In contrast to holistic IAA, these approaches provide more interpretable feedback by assessing specific visual attributes of images. In this line of work, several studies try to explicitly model visual attributes to provide more interpretable and fine-grained aesthetic assessment. For instance, Liu *et al.* proposed a composition-aware aesthetic assessment network, RGNet, which models the influence of composition on aesthetic perception by integrating image structure and regional contrast relationships (Liu et al. 2020). He *et al.* constructed a color aesthetics dataset, ICAA17K, and designed a color aesthetic assessment method based on interest points (He et al. 2023a, 2025a).

- Although existing IAA methods are categorized as holistic or multi-attribute, they are essentially similar, with the main differences lying in data-level design. Most of them are learning-based methods. We also draw on their design ideas and include them in our comparison experiments.
- Although existing IAA methods can be applied to the IC-TAA task, they typically assume aesthetic attributes are discrete and rely on heterogeneous images with fixed attribute values for learning, making them insensitive to continuously varying attributes such as color temperature and unable to reflect how aesthetic quality changes with other visual attributes under controlled content.
- To address the aforementioned issues, we propose IC-TAA, the task designed to assess the aesthetic influence of color temperature variations.

3 Methods

The proposed method involves two key components: approach and strategy. At the approach level, we design a multi-modal contrastive learning network, as shown in Fig. 2, that compares image pairs with different color temperatures by strictly controlling other attributes and incorporates textual features to predict aesthetic rankings. At the strategy level, we leverage color temperature transitivity, designing a weakly supervised strategy that samples images with different color temperature shifts rooted in anchor images and human perceptivity, establishes cross color temperature contrastive relations, and enables model learning from discrete labels. Additionally, we propose a new metric, IEA, to evaluate model performance across samples of varying difficulty as a complement to existing evaluation metrics.

3.1 Multimodal Contrastive Learning Framework

As shown in Fig. 2, the proposed framework consists of four components: a Color Temperature Encoder (CTE) for extracting color temperature features, a Contextual Aware Module (CAM) for contextual perception, a Cross-Modal Fusion Module (CFM) for multimodal interaction, and a Prediction and Ranking Predictor (PRP) for predicting the color temperature aesthetic ranking results.

Color Temperature Encoder. Considering the training cost, we froze the backbone of DINO and CLIP. To effectively extract image color temperature features, we design a CTE extraly, as shown in Fig. 2. The input image pair is first processed through convolutional blocks to extract low-level visual features, providing foundational information for color temperature analysis (Cheng et al. 2023). Considering that low-level features contain substantial noise unrelated to color temperature, we additionally introduce a lightweight attention module to enhance the robustness and relevance of the extracted color temperature features (Woo et al. 2018). Subsequently, the feature maps are processed by two sets of convolutional and normalization layers to further refine informative features related to color temperature. Finally, global average pooling is applied to aggregate the feature maps, generating an overall perceptual feature of color temperature F_c that serves as a fundamental support for subsequent aesthetic evaluation.

Contextual Awareness Module To enhance model understanding of image content and themes, the CAM is designed to integrate both visual and textual information. As shown in Fig. 2, the module consists of three components: a Visual

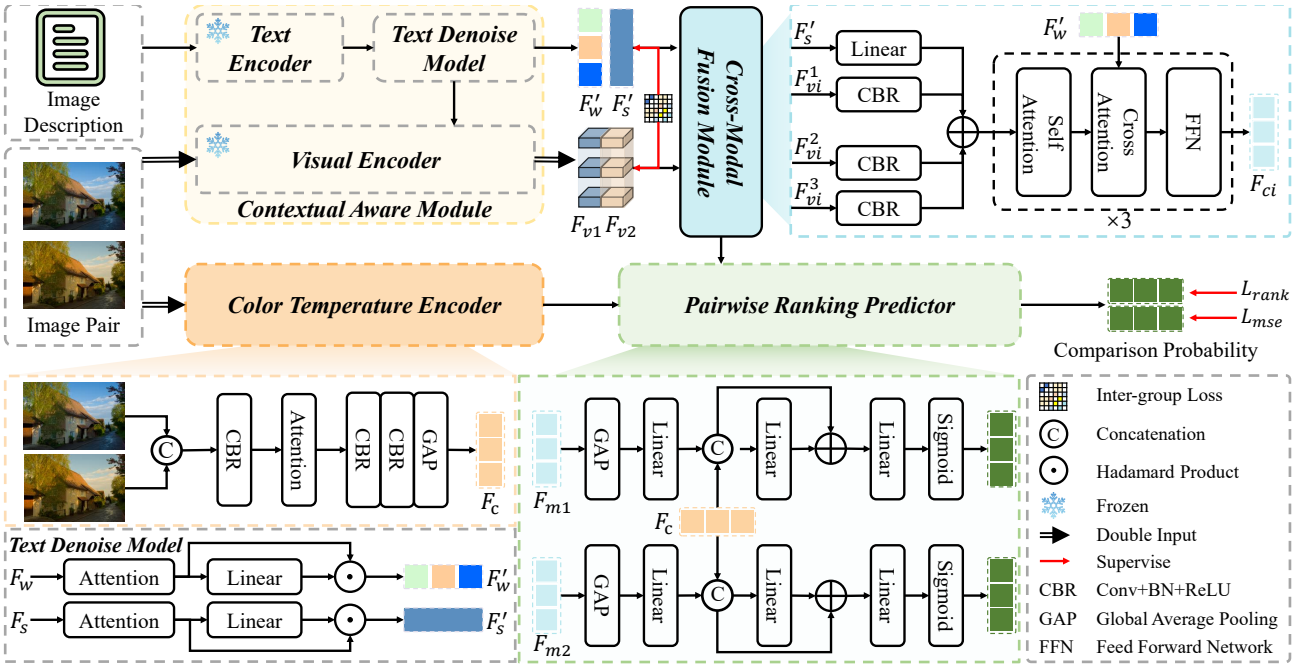

Figure 2: Overall framework of ICTA2Net, comprising four components: a Color Temperature Encoder for capturing color temperature variations; a Contextual Awareness Module (including Visual Encoder, Text Encoder, and Text Denoise Model); a Cross-Modal Fusion Module for visual-textual integration; and a Pairwise Ranking Predictor for aesthetic preference estimation.

Encoder (VE), a Text Encoder (TE), and a Text Denoising Module (TDM).

For visual encoding, we adopt the DINOv2, pretrained on ViT-B/14 (Oquab et al. 2023) as the VE. Given an image pair I_i, $i \in \{1, 2\}$, combined with textual prompts, we extract hierarchical visual features from different layers, selecting two intermediate layers and the final output to construct the visual representation set of the image, denoted as F_{vi}^j, where $i \in \{1, 2\}$ represents the two input images and $j \in \{1, 2, 3\}$ indicates the feature layer index.

For text encoding, we use the CLIP, based on the Transformer architecture and pre-trained on large-scale image-text pairs as TE, to extract word-level features F_w and sentence-level features F_s (Radford et al. 2021). Considering that textual inputs generally contain redundant or noisy information, we introduce a lightweight TDM to refine the text features. As shown in Fig. 2, the TDM takes F_w and F_s as input, models their contextual dependencies via a multi-head self-attention mechanism, and fuses them through linear transformation and element-wise multiplication to obtain denoised features F'_w and F'_s. This module effectively suppresses semantically irrelevant noise while preserving core semantic information. Finally, the denoised text features are fed into the VE to enable more accurate cross-modal semantic alignment. The effectiveness of this design is demonstrated in the ablation study, as shown in Table 4.

Cross-modal Fusion Module. To enhance model understanding of the relationship between image color temperature and contextual semantics, we propose a CFM, as illustrated in Fig. 2. Specifically, the multi-level visual features F_{vi}^j extracted from the VE and the sentence-level feature F'_s extracted from the TDM are first projected into a unified feature space through convolution and linear transformation, respectively. Then, the sentence-level feature is added with the three layers of visual features to produce the initially fused multimodal features F_{ci}, $i \in \{1, 2\}$. To further model semantic dependencies across modalities, we introduce a multi-layer attention interaction structure, which consists of alternately stacked Multi-Head Self-Attention (MHSA) and Multi-Head Cross-Attention (MHCA) blocks to strengthen intra-modal and inter-modal feature interactions. Specifically, each F_{ci} first passes through MHSA to capture both local and global dependencies within itself. Then, using the visual feature as the query and the word-level text features F'_w as the key and value, MHCA enables semantic alignment and complementary fusion across modalities. The final output is the fused feature representation F_{mi}, $i \in \{1, 2\}$.

Pairwise Ranking Predictor. After obtaining the color temperature perception features F_c and the visual-text fused features F_{mi}, we employ the PRP to perform relative aesthetic comparison between the image pair, as illustrated in Fig. 2. Specifically, for each image, the color temperature feature F_c and the semantic fusion feature F_{mi} are respectively processed via global pooling and dimensionality reduction. The resulting features are then concatenated and passed through a linear transformation followed by a residual connection to obtain the enhanced feature representation F_{ei}. Finally, a linear mapping followed by a Sigmoid is applied to produce a relative aesthetic score.

3.2 Weakly Supervised Strategy

As discussed in the *Introduction*, the human visual system exhibits the JND characteristic to changes in image color, When the color temperature shifts in a consistent direction, the corresponding aesthetic perception of image demonstrates a degree of transitivity, the observation was validated through a subjective experiment (please refer to Appendix A.1). Based on these thinking, we propose a structured weakly supervised strategy. Specifically, taking an anchor image (GT) as a reference, each group of images is discretely sampled in both the cool-shift and warm-shift directions based on the perceptual sensitivity of human vision. Then, the sampled images are paired with the anchor image as well as with other sampled images having the same color temperature shift direction, to construct contrastive relationships among images with different color temperature shifts, and effective supervision signals can be provided without massive, precise annotations. For further details, please refer to Appendix A.2.

3.3 IEA Metric

Human preferences for color temperature exhibit a certain degree of subjectivity, so the labeling results of color temperature usually present a preference distribution. Since ICTAA ultimately makes decisions based on preferences, the evaluation metric should take into account the intensity of such preferences. To this end, we propose the IEA, which weights the accuracy rate according to the confidence level of human preferences to reflect the model's performance under different sample difficulty levels, thereby supplementing the existing evaluation metrics.

IEA Definition. Given an image pair A and B, let $Vote_A$ and $Vote_B$ denote the number of expert preference votes received by images A and B, respectively. We first define the probability of the majority preference as:

$$P = \frac{max(Vote_A, Vote_B)}{Vote_A + Vote_B}. \tag{1}$$

Subsequently, the uncertainty of a sample is quantified based on the information entropy of the voting distribution:

$$H(p) = \begin{cases} 1, & \text{if } p = 1, \\ -p \cdot \log_2(p) - (1-p) \cdot \log_2(1-p), & \text{if } 0.5 \le p \le 1. \end{cases} \tag{2}$$

Accordingly, after normalization, the sample confidence is computed as:

$$c = 1 - 0.5 \cdot H(P). \tag{3}$$

The confidence score c represents the preference consistency of annotators, with higher values indicating greater consistency. The introduced IEA can better reflect the model's accuracy across samples of varying difficulty,

$$IEA = \frac{\sum_i c_i \cdot \mathbb{I}\left[y_i = \hat{y}_i\right]}{\sum_i c_i}, \tag{4}$$

where y_i denotes the ground truth label determined by majority voting of experts for the i-th sample, $\hat{y}_i$ represents the model prediction, c_i is the corresponding confidence of the sample, $\mathbb{I}[]$ is an indicator function, returning 1 if the condition holds true, and 0 otherwise.

3.4 Loss Functions

Inter-group Contrastive Loss. To improve model performance in cross-modal semantic alignment, we introduce an inter-group contrastive loss. Specifically, positive samples are constructed within the same group, while negative samples are constructed within different groups. Contrastive learning is then applied to pull intra-group positives closer and push inter-group negatives farther apart. The loss is defined as follows:

$$\mathcal{L} = \frac{1}{2B} \sum_{i=1}^{2B} -\log\left(\frac{N_i}{D_i + \varepsilon}\right), \tag{5}$$

$$\begin{aligned} N_i &= \sum_{\substack{j=1 \\ j \neq i}}^{2B} \mathbb{I}[g_i = g_j] \cdot \exp\left(\frac{\hat{\mathbf{v}}_i^\top \hat{\mathbf{t}}_j}{\tau}\right), \\ D_i &= \sum_{j=1}^{2B} \exp\left(\frac{\hat{\mathbf{v}}_i^\top \hat{\mathbf{t}}_j}{\tau}\right) - \exp\left(\frac{\hat{\mathbf{v}}_i^\top \hat{\mathbf{t}}_i}{\tau}\right). \end{aligned} \tag{6}$$

Here, B denotes the batch size, $\hat{\mathbf{v}}_i$ represents the i-th normalized visual feature vector, and $\hat{\mathbf{t}}_i$ denotes the j-th normalized textual feature vector. τ is the temperature scaling parameter. g_i indicates the group label of the i-th sample. $\mathbb{I}[\cdot]$ is an indicator function that returns 1 when the condition is true and 0 otherwise, and the ε is a small constant added for numerical stability.

Total Loss. To simultaneously improve perceptual accuracy in color temperature aesthetics, relative ranking capability, and contextual discriminability of the model, we employ a total of three loss functions: the mean squared error (MSE) loss, the ranking loss (Kong et al. 2016), and inter-group loss. The overall training loss is defined as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{mse} + \lambda_2 \mathcal{L}_{rank} + \lambda_3 \mathcal{L}_{inter}. \tag{7}$$

Among them, λ_1, λ_2, and λ_3 are weighting coefficients used to balance the contributions of each loss function.

4 Dataset Construction

4.1 Image Collection and Generation

To train our model, we require numerous images with realistic cool and warm color shifts and their corresponding aesthetic ground truths. To this end, we construct the dataset, ICTAA240K, which is built upon two high-quality public RAW image datasets, MIT-Adobe FiveK (Bychkovsky et al. 2011) and PPR10K (Liang et al. 2021), both of which were rendered by expert photographers. From these datasets, we selected a total of 5,748 images covering diverse subjects, photographic styles, and common color temperature ranges. For each RAW image, we use the Adobe Camera Raw SDK (Adobe Inc. 2024) to simulate camera-applied color temperatures, which accurately emulate the nonlinear rendering process based on metadata embedded in each DNG file (Afifi et al. 2019; Ronneberger, Fischer, and Brox 2015). We render each RAW RGB image with a different color temperature to mimic a real color temperature shift.

Specifically, for each image, the color temperature is adjusted in 200K steps from 2000K to 10000K, accounting for human perceptual sensitivity. The ground-truth (GT) images are photographs manually retouched by professional photographers (Expert C in (Bychkovsky et al. 2011) and Expert A in (Liang et al. 2021)), which serve as high-quality aesthetic references. In total, we generate 241,416 sRGB images covering a representative spectrum from cool to warm tones (mean color difference $\Delta E_{2000} = 0.69$, which is less than 1, indicating an extremely small color variation). To further support multimodal learning and aesthetic preference modeling, we provide textual descriptions for each image, covering three key dimensions: theme, content, and photographic category. The textual annotations are generated using GPT-4o (Hurst et al. 2024), and refined through human feedback to ensure accuracy. The details for dataset construction pipeline, color temperature distribution, and color difference analysis are illustrated in Appendix A.3.

4.2 Test Dataset Construction

To evaluate the model's ability to capture aesthetic preferences for image color temperature, 20% of the FiveK dataset and all portrait images from the PPR10K dataset are used for testing. Considering manual costs, we sample 8 representative images from each group based on the perceptual sensitivity of human vision and the color difference between each shift image and the anchor image (GT), covering typical color temperature deviations ranging from noticeably cool to warm. Then, the selected images are divided into general scene images and portrait scene images. We ultimately construct four test datasets, as shown in Table 1. 1) ICTAA-GP / ICTAA-HP: Constructed for general and portrait scenes, respectively. Each image pair consists of a cool-shift and a warm-shift version and is annotated with pairwise preference labels by 10 annotators with backgrounds in aesthetic research. The preference confidence scores are then computed based on the annotators' voting results. 2) ICTAA-GF / ICTAA-HF: Extend the above test datasets by introducing an anchor image (GT) for each pair, enabling evaluation of whether the model can learn aesthetic trends relative to the anchor image. During annotation, we strictly control for confounding variables other than color temperature, ensuring that preference judgments are solely influenced by color temperature differences. The annotation process and quality control mechanisms are detailed in Appendix A.3.

Subset Name	Number	Type	Scene
ICTAA-GP	1,142	C vs. W	General Scenes
ICTAA-HP	1,592	C vs. W	Portrait Scenes
ICTAA-GF	4,628	C vs. W, C/W vs GT	General Scenes
ICTAA-HF	6,905	C vs. W, C/W vs GT	Portrait Scenes

Table 1. Details of ICTAA Dataset Subsets. It contains four subsets, where C represents Cool-shift, W represents Warm-shift, and GT represents images adjusted by experts.

5 Experiments

5.1 Implementation Details

We implemented ICTA2Net using the PyTorch framework and trained the model with the Adam optimizer. During training, the batch size was set to 24, with an input image resolution of 224×224 pixels. The initial learning rate was set to 2×10^{-5} and dynamically adjusted using a cosine annealing schedule. The training process was conducted for 10 epochs. All experiments were performed on a single NVIDIA GeForce RTX 3090 GPU.

5.2 Benchmark Datasets and Models

The model was trained on the ICTAA240K training dataset and evaluated on four designated test subsets: ICTAA-GP, ICTAA-GF, ICTAA-HP, and ICTAA-HF.

To the best of our knowledge, there are no existing methods specifically designed for ICTAA. Therefore, we selected ten related representative IAA models- NIMA (Talebi and Milanfar 2018), BLG-PIAA (Zhu et al. 2020), HGCN (She et al. 2021), MaxViT (Tu et al. 2022), TANet (He et al. 2022), EAT (He et al. 2023b), DeT (He et al. 2023a), BAID (Yi et al. 2023), EAMBNet (Chen et al. 2024), and Prompt-DeT (He et al. 2025b). All of these models are selected based on the following two criteria: 1) availability of public implementation code; 2) proven representativeness or strong performance in aesthetic quality assessment tasks. To ensure a fair comparison, these models were retrained in our dataset using their publicly available training configurations. Furthermore, to align with the contrastive learning framework proposed in this study, the baseline models were integrated into our contrastive learning training pipeline without altering their original network architectures. The specific process is detailed in Appendix A.4.

5.3 Experimental Results and Analysis

Quantitative Analysis. To comprehensively evaluate the preference of the proposed method, we adopt two metrics, IEA and ACC, comparing our methods against ten SOTA models. As shown in Table 2, ICTA2Net achieves highly competitive performance across two metrics of four test datasets, demonstrating its effectiveness and superiority in the ICTAA task. Notably, on the ICTAA-HP test dataset, ICTA2Net significantly outperforms all baseline methods, improving by 9.3% and 8.2%, respectively, compared with the second place in the two metrics of ACC and IEA, indicating its stronger generalization ability and robustness when facing unseen samples.

Visualization Analysis. Fig. 11 shows the t-SNE visualization of the output features from ICTA2Net (please refer to Appendix A.5), which verifies that the proposed model can effectively extract color temperature and semantic information. Figs. 13, 14, 15, and 16 show the ranking results of image color temperature aesthetics by our proposed method, which are well consistent with human perception. Please refer to Appendix A.6.

5.4 IEA Validity Verification

Table 5 shows that for all baselines, the Spearman (SRCC) and Pearson (PLCC) correlation coefficients between ACC and IEA exceed 0.98, indicating highly consistent evaluation trends, confirming the reliability of the proposed IEA metric. To further verify its effectiveness, samples are divided

Model	Year	Publish	ICTAA-GP		ICTAA-GF		ICTAA-HP		ICTAA-HF	
			ACC↑	IEA↑	ACC↑	IEA↑	ACC↑	IEA↑	ACC↑	IEA↑
NIMA	2018	TIP	0.630	0.687	0.774	0.795	0.569	0.582	0.754	0.776
BLG-PIAA	2020	TCYB	0.659	0.706	0.846	0.871	0.616	0.656	0.822	0.847
HGCN	2021	CVPR	0.673	0.726	0.806	0.828	0.658	0.705	0.799	0.819
MaxViT	2022	ECCV	0.685	0.733	0.855	0.880	0.664	0.717	0.838	0.861
TANet	2022	IJCAI	0.683	0.735	0.828	0.854	0.634	0.679	0.813	0.832
EAT	2023	ACM MM	0.696	0.746	0.832	0.852	0.653	0.703	0.823	0.845
DeT	2023	ICCV	0.483	0.467	0.533	0.543	0.438	0.457	0.501	0.508
BAID	2023	CVPR	0.664	0.710	0.839	0.863	0.631	0.674	0.821	0.845
EAMBNet	2024	TMM	0.646	0.684	0.741	0.764	0.533	0.563	0.773	0.793
Prompt_DeT	2025	INF. FUSION	0.652	0.698	0.836	0.861	0.647	0.687	0.828	0.851
Ours			0.700	0.746	0.870	0.894	0.726	0.776	0.875	0.896

Table 2. Comparisons of 10 models on our ICTAA240K dataset. To ensure the fairness of the comparison, all models adopted a unified experimental paradigm: they were strictly retrained and tested on this dataset following the parameters recommended in their original papers. ↑ indicates that the larger the corresponding metric value, the better.

into high- and low-divergence groups to assess model performance under varying difficulty (please see Appendix A.7 for more detail).

Metric	Dataset			
	ICTAA-GP	ICTAA-GF	ICTAA-HP	ICTAA-HF
SRCC	0.98	0.99	1.00	0.98
PLCC	0.99	0.99	0.99	0.99

Table 3. IEA Validation Results.

5.5 Ablation Study

To assess the contribution of each component to the overall performance of ICTA2Net, we conducted ablation studies by retraining the model with specific modules removed or replaced. The results on the ICTAA-GP, ICTAA-GF, ICTAA-HP, and ICTAA-HF datasets are shown in Table 4.

#	Ablation Model	**IEA ↑**			
		GP	GF	HP	HF
1	w/o. Loss	0.744	0.891	0.761	0.892
2	w/o. Text	0.746	0.891	0.759	0.892
3	w/o. CTE	0.712	0.885	0.649	0.874
4	w/o. TDM	0.751	0.892	0.744	0.891
5	w/o. CFM	0.754	0.896	0.710	0.875
6	CTA^{2}Net	0.746	0.894	0.776	0.896

Table 4. Ablation Study Results. (GP: ICTAA-GP, GF: ICTAA-GF, HP: ICTAA-HP, HF: ICTAA-HF)

From the presented results, we have several observations. 1) Comparison between #1 and #6 shows that removing the inter-group contrastive loss leads to performance degradation on all datasets, especially on the ICTAA-HP dataset. This highlights the effectiveness of the inter-group loss in promoting cross-modal semantic alignment, thereby improving its aesthetic computation capability. 2) Comparison between #2 and #6 shows that without textual information, results in a certain degree of performance degradation, which is particularly evident on the ICTAA-HP dataset. This indicates that combining textual and visual information enables a more accurate understanding of color temperature aesthetic expression. 3) Comparison between #3 and #6 shows a significant performance drop across all datasets, highlighting the important role of CTE in capturing the relationship between color temperature variation and aesthetic perception. Moreover, 4) Replacing the TDM with a simple linear connection (comparison between #4 and #6) results in a noticeable performance drop on ICTAA-HP and fluctuations on other datasets. This indicates that the TDM contributes to noise suppression and feature enhancement. 5) Replacing the CFM module with a simple concatenation operation (comparison between #5 and #6) leads to substantial performance drops on both ICTAA-HP and ICTAA-HF, validating the importance of our proposed CFM in effectively aligning textual and visual information.

In summary, the experimental results demonstrate the necessity of all proposed components for the ICTAA task, with the inter-group contrastive loss, CTN, and CFM playing particularly critical roles. Furthermore, the significant performance degradation observed on the ICTAA-HP dataset after removing any component further confirms the indispensable importance of these modules in improving cross-modal semantic alignment of the model.

6 Conclusion

In this work, we first propose and define the ICTAA task. To address its key challenges, perception sensitivity and spectral continuity, we develop the pairwise contrastive learning framework ICTA2Net with a structured weakly supervised strategy guided by anchor samples. Furthermore, we introduce an IEA accuracy metric based on information entropy-weighted, and construct a large-scale dataset. Our benchmark validation confirms the superiority of the method.

However, this work has several limitations. First, by focusing solely on color temperature, the evaluation of white balance effects remains incomplete. Second, the weakly supervised strategy adopted to avoid the constraints of spectral continuity is not an optimal solution. Future work will expand the assessment framework to include multidimensional attributes such as hue and saturation, and to refine the supervision mechanism in pursuit of more effective solutions.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 62502040, the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation under Grant GZC20251056 and GZC20251061, Science and Technology Research Project of Henan Province 252102211037. Furthermore, we would like to express our special gratitude to Xiang Liu, Photographer and General Manager of Visual China Group 500px, for his guidance on this work.

References

Adobe Inc. 2024. Adjust color in Camera Raw – Photoshop Elements. Available at: https://helpx.adobe.com/photoshop-elements/using/color-camera-raw.html. Accessed: 2024-12-01. 5

Afifi, M.; Punnappurath, A.; Abdelhamed, A.; Karaimer, H. C.; Abuolaim, A.; and Brown, M. S. 2019. Color temperature tuning: Allowing accurate post-capture white-balance editing. In *Color and Imaging Conference (CIC)*. 5

Afifi, M.; Zhao, L.; Punnappurath, A.; Abdelsalam, M. A.; Zhang, R.; and Brown, M. S. 2025. Time-Aware Auto White Balance in Mobile Photography. In *The IEEE International Conference on Computer Vision (ICCV)*. 2, 3

Barron, J. T. 2015. Convolutional color constancy. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2

Bianco, S.; and Cusano, C. 2019. Quasi-unsupervised color constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

Bychkovsky, V.; Paris, S.; Chan, E.; and Durand, F. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 5, 6

Chen, H.; Shao, F.; Mu, B.; and Jiang, Q. 2024. Image aesthetics assessment with emotion-aware multibranch network. *IEEE Transactions on Instrumentation and Measurement (TIM)*, 73: 1–15. 6

Cheng, J.; Wu, Z.; Wang, S.; Demonceaux, C.; and Jiang, Q. 2023. Bidirectional collaborative mentoring network for marine organism detection and beyond. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 33: 6595–6608. 3

DXOMARK. 2024. DXOMARK Analyzer Catalogue 2024. https://corp.dxomark.com/wp-content/uploads/2024/06/DXOMARK_AnalyzerCatalogue_2024.pdf. 1

He, S.; Ming, A.; Li, Y.; Sun, J.; Zheng, S.; and Ma, H. 2023a. Thinking image color aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3, 6

He, S.; Ming, A.; Zheng, S.; Zhong, H.; and Ma, H. 2023b. Eat: An enhancer for aesthetics-oriented transformers. In *Proceedings of the 31st ACM international conference on multimedia (ACM MM)*. 6

He, S.; Xiao, Y.; Ming, A.; and Ma, H. 2025a. Prompt-guided image color aesthetics assessment: Models, datasets and benchmarks. *Information Fusion*, 114: 102706. 3

He, S.; Xiao, Y.; Ming, A.; and Ma, H. 2025b. Prompt-guided image color aesthetics assessment: Models, datasets and benchmarks. *Information Fusion (NFORM FUSION)*, 114: 102706. 6

He, S.; Zhang, Y.; Xie, R.; Jiang, D.; and Ming, A. 2022. Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*. 1, 6

Hosu, V.; Goldlucke, B.; and Saupe, D. 2019. Effective aesthetics prediction with multi-level spatially pooled features. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 3

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*. 6

Jiang, A.; Yao, X.; Westland, S.; Hemingray, C.; Foing, B.; and Lin, J. 2022. The effect of correlated colour temperature on physiological, emotional and subjective satisfaction in the hygiene area of a space station. *International Journal of Environmental Research and Public Health (IJERPH)*, 19: 9090. 1

Jiang, Q.; Liu, F.; Wang, Z.; Wang, S.; and Lin, W. 2024. Rethinking and Conceptualizing Just Noticeable Difference Estimation by Residual Learning. *IEEE Transactions on Circuits and Systems for Video Technology (TSCVT)*, 34: 9515–9527. 2

Jin, X.; Wu, L.; Zhao, G.; Li, X.; Zhang, X.; Ge, S.; Zou, D.; Zhou, B.; and Zhou, X. 2019. Aesthetic attributes assessment of images. In *Proceedings of the 27th ACM international conference on multimedia (ACM MM)*. 3

Kong, S.; Shen, X.; Lin, Z.; Mech, R.; and Fowlkes, C. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *Computer Vision–ECCV 2016: 14th European Conference (ECCV)*. 3, 5

Kınlı, F.; Yılmaz, D.; Özcan, B.; and Kıraç, F. 2023. Modeling the Lighting in Scenes as Style for Auto White-Balance Correction. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2

Lam, E. Y. 2005. Combining gray world and retinex theory for automatic white balance in digital photography. In *Proceedings of the Ninth International Symposium on Consumer Electronics, 2005.(ISCE)*. 2

Leder, H.; Hakala, J.; Peltoketo, V.-T.; Valuch, C.; and Pelowski, M. 2022. Swipes and saves: A taxonomy of factors influencing aesthetic assessments and perceived beauty of mobile phone photographs. *Frontiers in psychology (FRONT PSYCHOL)*, 13: 786977. 1

Li, C.; Kang, X.; and Ming, A. 2023. WBFlow: Few-shot white balance for sRGB images via reversible neural flows. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 2

Li, C.; Kang, X.; Zhang, Z.; and Ming, A. 2023. SWBNet: a stable white balance network for sRGB images. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2

Li, Y.; Ru, T.; Chen, Q.; Qian, L.; Luo, X.; and Zhou, G. 2021. Effects of illuminance and correlated color temperature of indoor light on emotion perception. *Scientific reports (SCI REP)*, 11: 14351. 1

Liang, J.; Zeng, H.; Cui, M.; Xie, X.; and Zhang, L. 2021. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5, 6

Liu, D.; Puri, R.; Kamath, N.; and Bhattacharya, S. 2020. Composition-aware image aesthetics assessment. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*. 3

Liu, L.; He, S.; Ming, A.; Xie, R.; and Ma, H. 2024. ELTA: an enhancer against long-tail for aesthetics-oriented models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. 2

Liu, Y.-C.; Chan, W.-H.; and Chen, Y.-Q. 1995. Automatic white balance for digital still camera. *IEEE Transactions on Consumer Electronics (TCE)*, 41: 460–466. 2

MacAdam, D. L. 1942. Visual sensitivities to color differences in daylight. *Journal of the Optical Society of America (J. Opt. Soc. Am.)*, 32: 247–274. 2

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Howes, R.; Huang, P.-Y.; Xu, H.; Sharma, V.; Li, S.-W.; Galuba, W.; Rabbat, M.; Assran, M.; Ballas, N.; Synnaeve, G.; Misra, I.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision. 4

Ou, L.-C.; Luo, M. R.; Woodcock, A.; and Wright, A. 2004. A study of colour emotion and colour preference. Part I: Colour emotions for single colours. *Color Research & Application (COLOR RES APPL)*, 29: 232–240. 1

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*. 4

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*. 5

Sharma, G.; Wu, W.; and Dalal, E. 2005. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application (COLOR RES APPL)*, 30: 21 – 30. 2

She, D.; Lai, Y.-K.; Yi, G.; and Xu, K. 2021. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 6

Sheng, K.; Dong, W.; Ma, C.; Mei, X.; Huang, F.; and Hu, B.-G. 2018. Attention-based multi-patch aggregation for image aesthetic assessment. In *Proceedings of the 26th ACM international conference on Multimedia (ACM MM)*. 3

Talebi, H.; and Milanfar, P. 2018. NIMA: Neural image assessment. *IEEE transactions on image processing (TIP)*, 27: 3998–4011. 1, 3, 6

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. Maxvit: Multi-axis vision transformer. In *European conference on computer vision (ECCV)*. 6

Valdez, P.; and Mehrabian, A. 1994. Effects of color on emotions. *Journal of experimental psychology: General (J EXP PSYCHOL GEN)*, 123: 394. 1

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; Yan, Q.; Min, X.; Zhai, G.; and Lin, W. 2024. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. 3

Xie, R.; Ming, A.; He, S.; Xiao, Y.; and Ma, H. 2024. " Special Relativity" of Image Aesthetics Assessment: a Preliminary Empirical Perspective. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 3

Yi, R.; Tian, H.; Gu, Z.; Lai, Y.-K.; and Rosin, P. L. 2023. Towards Artistic Image Aesthetics Assessment: A Large-Scale Dataset and a New Method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1, 6

Zhu, H.; Li, L.; Wu, J.; Zhao, S.; Ding, G.; and Shi, G. 2020. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *IEEE Transactions on Cybernetics (TCYB)*, 52: 1798–1811. 6