Supplementary Materials

A APPENDIX

A.1 Elucidation for the Foundations of SR-IAA

I. Philosophical foundation.

1) "***Every judgment is by its form one-sided and, to that extent, false,***" stated by Hegel in Section 31 of the *Shorter Logic* [1].

2) ***The quantification methods need to accomplish two purposes, yet the existing approaches only fulfill one of them.*** The prevailing methods of quantification involve scoring images, which serves two essential purposes: positioning a sample within the entire sample space and assessing aesthetic differences between samples. The existing methods, however, primarily calculate positioning and then attribute the disparity in positioning to variations in image aesthetics. Nevertheless, this inference is likely incorrect, just as the spatial distance of colors in RGB space does not align with differences in human perception.

3) ***SR-IAA fully considers and achieves both purposes.*** SR-IAA calculates positioning based on relative aesthetic preferences between pairs of images, while acknowledging that the difference in positioning may not necessarily correspond to the difference in image aesthetics. Instead, it assumes that different aesthetic values define the distinction in aesthetics between two images.

In summary, ***SR-IAA exhibits a higher level of philosophical underpinnings compared to existing quantization methods due to its comprehensive nature.***

II. Psychological foundation.

1) ***The "stimulus-respons" mechanism and the instinct of "avoidance of harm"***are fundamental biological responses, which have been substantiated by scientific research in biology.

2) ***The aforementioned biological instinct is manifested in human beings as the cognitive construct of "gain and loss."*** The research on prosocial behavior by Feng *et al.* [2] states, "People treat losses and gains differently during individual decision-making, which is closely related to differential emotional responses to losses and gains."

3) ***The perception of attractiveness and unattractiveness is a psychological instinct inherent in human beings.*** Psychologically, individuals tend to gain greater solace in aesthetically pleasing imagery, while losing an elevated sense of comfort when exposed to visually unappealing stimuli. Mo *et al.* [3] also demonstrated that "humans automatically exhibit preference for visual and moral beauty without explicit cognitive efforts."

In summary, ***the human capacity to discriminate between two comparable images and determine their relative aesthetic appeal or unattractiveness can be attributed to an inherent psychological instinct.***

III. Mathematical foundation.

1) ***The image aesthetics does not conform to a total order relation [4], and current methods for aesthetics assessment based on image scoring are almost invariably subject to the "Condorcet Paradox" [5].*** A total order relation implies a sequential relationship among all elements within a given set, manifesting in the aesthetics domain as a clear judgement of beauty or ugliness between any two images. However, a simple example suffices to demonstrate that the aesthetic comparability of any two images cannot be assured: there is no feasible method to compare the themes of probability statistics and natural scenery. Existing approaches to aesthetics assessment entail direct scoring of images, predicated on the erroneous assumption that a higher score invariably indicates superior aesthetic value compared to a lower score. Furthermore, the aggregation of dataset annotations from multiple annotators' collective votes introduces the "Voting Paradox" [1], casting doubt on the scientific validity of the quantification results and the dataset itself.

2) ***Image aesthetics neither satisfies a total order nor a partial order relation.*** Both partial and total order relation embody the concept of "order," characterized by reflexivity, anti-symmetry, and transitivity. A total order relation suggests an established order among all elements within a given set and can be considered a special case of a partial order relation. Since aesthetics does not ensure comparability between any two images, it fails to fulfill the criteria for a total order relation. Moreover, because aesthetic judgment may not adhere to transitivity, it also does not meet the requirements of a partial order relation.

3) ***Images sharing the same theme can be approximately considered to satisfy a total order relation [4], whereas collections of images from different themes can be thought of as exhibiting a partial order relation [4].*** To utilize existing datasets, which are predicated on the assumption of a total order relationship, it becomes necessary to adopt the aforementioned assumptions. Subsequently, we embark on methodological design and experimental validation, culminating in the achievement of SOTA performance.

In summary, existing methods are based on the total order of image aesthetic data, and the utilization of a multi-voter mechanism can result in the "Condorcet Paradox," raising suspicions quantification results. On the other hand, SR-IAA argues that image aesthetic relationships do not adhere to even a partial order. However, in order to leverage existing data, a compromise can be made based on thematic considerations, which is considerably more reasonable compared to existing methods.

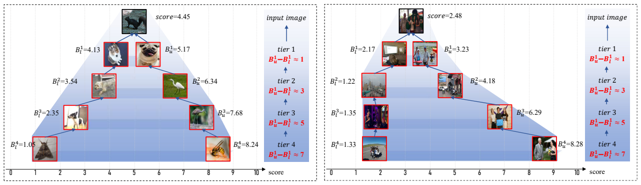
A.2 Reliability and Accuracy of Triplet Data in Four-tier Training for RSM

Under all circumstances, the reference images can reliably and accurately constrain the input images through a meticulously designed data construction method, denoted as Algorithm 1. In Fig. 6, we present examples of triplet data based on Algorithm 1 under two different scenarios. Fig. 6 (a) demonstrates the situation where both B_l^e and B_u^e can expand normally as the tier increases; Fig. 6 (b) illustrates the case where the left boundary B_l^e can no longer expand, resulting in an increased expansion speed for the other boundary

---

[1] The 18th-century French philosopher Condorcet proposed the famous "Voting Paradox," also known as the "Condorcet Paradox": Suppose there are three people, voter1, voter2 and voter3, facing three alternative options, ABC, with the following preference rankings: voter1 prefers A>B>C; voter2 prefers B>C>A; voter3 prefers C>A>B. Since both voter1 and voter2 think B is better than C, according to the principle of minority obedience to the majority, society should also think B is better than C; similarly, both voter2 and voter3 think C is better than A, society should also think C is better than A. So society thinks B is better than A. However, both voter1 and voter3 think A is better than B, so a contradiction arises.

B_u^e. Regardless of the initial position of the input image, we ensure that the difference between B_l^1 and B_u^1 is close to 1 in the final stage (tier= 1) of training, thereby ensuring the robust performance of the RSM.



(a) Four-tier Training Example 1

(a) Four-tier Training Example 2

Figure 6: Examples of Four-tier Training for RSM.

A.3 Inevitable but not Substantial Computational Cost due to Our Comparison Strategy

Our approach significantly enhances the performance of IAA tasks, howbeit at the expense of computational efficiency. This is attributed to our utilization of the triplet comparison strategy in PEM for concurrent processing of three images. In contrast to those single-image processing techniques, our approach incurs a relatively higher computational cost but remains within the same order of magnitude, as demonstrated in Table 4.

Method	Image size	Params	Throughput
Kong *et al.*	256×256	-	24.48 image/s
MP_{ada}	224×224	-	31.77 image/s
Malu *et al.*	299×299	23.5M	32.96 image/s
NIMA	224×224	14.7M	17.32 image/s
MUSIQ	768×1024	27.0M	16.94 image/s
TANet	224×224	13.8M	10.84 image/s
MaxViT	224×224	30.9M	16.54 image/s
EAT	224×224	87.0M	10.83 image/s
Ours	768×1024	28.5M	6.36 image/s

Table 4: Comparison of computational efficiencies among different methods, with throughput measured using a single RTX 3060 GPU.

A.4 Effectiveness of Operations with Different Intensity Levels

An image undergoes various intensity editing operations, resulting in different levels of aesthetic degradation while maintaining thematic consistency. Consequently, the aesthetic disparities between images can be roughly considered proportional to the intensity levels of image editing. Fig. 7 illustrates the image editing operations employed in our self-supervised learning approach, which induce aesthetic degradations to the input image. For images subjected to the same operation but with different intensity levels, they share the same theme, thereby roughly satisfying the transitivity described in Compromise I. Under these conditions, we can employ editing intensity as a measure of aesthetic variation, thus ensuring the rigor of our self-supervised training.

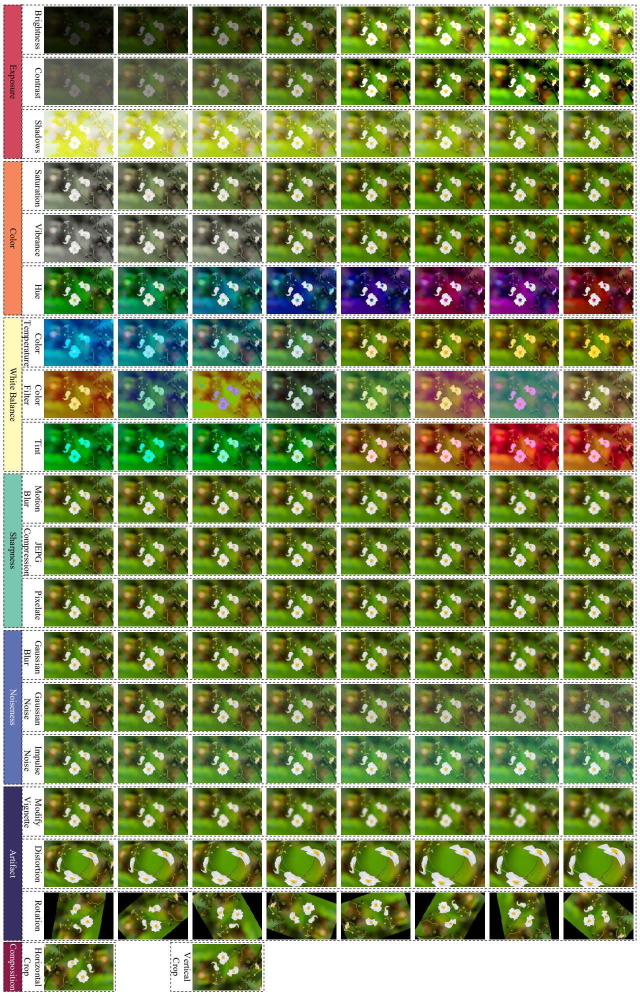


Figure 7: Visualization examples of our operations with different intensity levels.

REFERENCES

[1] J.G. Hibben and E. Luft. *Hegel's Shorter Logic: An Introduction and Commentary*. Gegensatz Press, 2013.

[2] Yijie Zhang Zhixin Zhang Jie Yuan Chunliang Feng, Chunliang Feng. Prosocial gains and losses: Modulations of human social decision-making by loss-gain context. *Frontiers in Psychology*, 2021.

[3] Kaixin Qin Lei Mo Ce Mo, Tiansheng Xia. Natural tendency towards beauty in humans: Evidence from binocular rivalry. *PLOS ONE*, 11(3):e0150147, 2016.

[4] David Gries and Fred B. Schneider. *A Logical Approach to Discrete Math*. Springer New York, New York, 1993.

[5] William V Gehrlein. Condorcet's paradox. *Theory and decision*, 15(2):161–197, 1983.