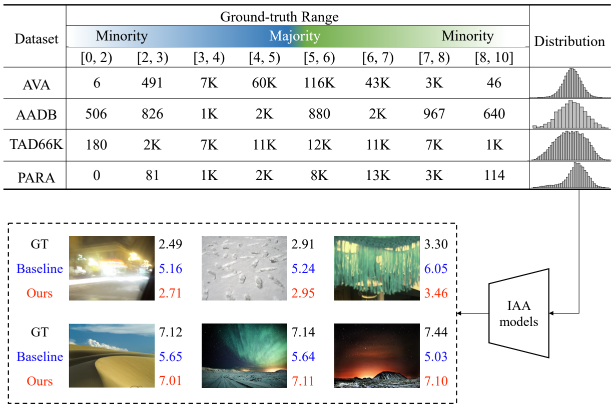
ELTA: An Enhancer against Long-Tail for Aesthetics-oriented Models

Limin Liu [*1] **Shuai He** [*1] **Anlong Ming** [*1] **Rui Xie** [1] **Huadong Ma** [1]

Abstract

Real-world datasets often exhibit long-tailed distributions, compromising the generalization and fairness of learning-based models. This issue is particularly pronounced in Image Aesthetics Assessment (IAA) tasks, where such imbalance is difficult to mitigate due to a severe distribution mismatch between features and labels, as well as the great sensitivity of aesthetics to image variations. To address these issues, we propose an Enhancer against Long-Tail for Aesthetics-oriented models (ELTA). ELTA first utilizes a dedicated mixup technique to enhance minority feature representation in high-level space while preserving their intrinsic aesthetic qualities. Next, it aligns features and labels through a similarity consistency approach, effectively alleviating the distribution mismatch. Finally, ELTA adopts a specific strategy to refine the output distribution, thereby enhancing the quality of pseudo-labels. Experiments on four representative datasets (AVA, AADB, TAD66K, and PARA) show that our proposed ELTA achieves state-of-the-art performance by effectively mitigating the long-tailed issue in IAA datasets. Moreover, ELTA is designed with plug-and-play capabilities for seamless integration with existing methods. To our knowledge, this is the first contribution in the IAA community addressing long-tail. All resources are available in here.



Dataset	Ground-truth Range								Distribution
	Minority			Majority			Minority		
	[0, 2)	[2, 3)	[3, 4)	[4, 5)	[5, 6)	[6, 7)	[7, 8)	[8, 10]	
AVA	6	491	7K	60K	116K	43K	3K	46	
AADB	506	826	1K	2K	880	2K	967	640	
TAD66K	180	2K	7K	11K	12K	11K	7K	1K	
PARA	0	81	1K	2K	8K	13K	3K	114	

Figure 1: The long-tailed distributions in mainstream IAA datasets and resulting model bias (in blue) are mitigated by our proposed ELTA (in red), which addresses the issue of insufficient differentiation in model scores.

1. Introduction

Benefiting from the growth and utilization of large-scale datasets, deep neural networks have achieved remarkable successes in computer vision tasks. However, these networks often face challenges in real-world applications due to data imbalance. It is commonly observed that models trained on long-tailed data are biased towards majority groups with abundant samples, and away from minority groups with fewer samples. The bias compromises the model's generalizability and fairness, and raises concerns about ethical implications.

In some classical vision tasks, such as image classification and instance segmentation, various solutions (Kang et al., 2019; Cui et al., 2019; Yang & Xu, 2020; Wei et al., 2021; Yu et al., 2022; Ahn et al., 2023) are proposed to mitigate the data imbalance. However, the progress of handling long-tailed datasets in Image Aesthetics Assessment (IAA) lags behind classical vision tasks. As a recently emerging research, IAA is becoming increasingly vital in various domains, such as computational photography, art design, and recommender systems. Unfortunately, as shown in Figure 1, IAA encounters a particularly severe long-tail issue, which is further complicated by its specificity.

Lack of well-defined categories. Compared to the long-tail issues encountered in classification tasks, the IAA tasks present unique challenges. The subjectivity in the labeling process and the ambiguity between different aesthetic criteria levels make it difficult to define clear categories. Due to creating artificial boundaries based on scores in IAA is both

[*] Equal contribution [1] School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing, China. Correspondence to: Anlong Ming <mal@bupt.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

ambiguous and not universally applicable, this absence of well-defined categories hinders the use of many class-based methods such as re-sampling (Chawla et al., 2002; Kang et al., 2019; Wang et al., 2019; Bai et al., 2023), re-weighting (Cui et al., 2019; Cao et al., 2019; Ren et al., 2020; Wang et al., 2021b; Du et al., 2023), and data augmentation (Kim et al., 2020; Wang et al., 2021a; Ahn et al., 2023; Perrett et al., 2023).

Mismatch between features and labels. On the one hand, the inherent subjectivity of IAA datasets may lead to "noise". For example, the same image may contain different opinions, all of which are treated as ground truth. This uncertainty significantly increases the learning difficulty of models and leads to severe mismatches between labels and learning features (Figure 2). On the other hand, image aesthetics tend to be sensitive to some augmentations. Traditional augmentation at the raw-pixel level (Kim et al., 2020; Park et al., 2022; Ahn et al., 2023) may alter images' aesthetics without changing labels, resulting in augmented images with incorrect labels.

Smoothness of predicted distributions. Recent semi-supervised learning methods (Berthelot et al., 2019; Sohn et al., 2020; Zhang et al., 2021) rely heavily on model confidence and specific threshold to identify reliable pseudo-labels. Unlike typical models, the output logits of IAA models frequently exhibit smoother output distributions, where confident and uncertain predictions blend seamlessly. Consequently, simply filtering out uncertain predictions by a high threshold becomes ineffective for generating accurate pseudo-labels in IAA tasks.

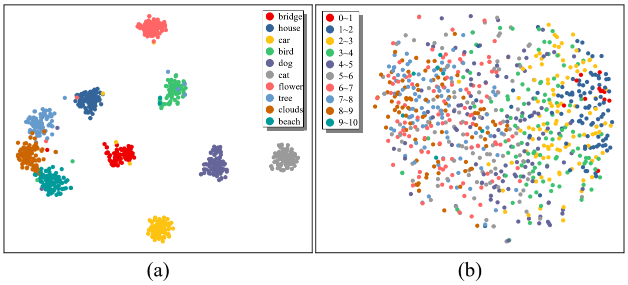


Figure 2: T-SNE visualization of feature distributions extracted by a classification network and an IAA network respectively. The entire score range in IAA is divided into ten segments, with feature points in each segment marked in different colors. The classification features (a) clearly distinguish between different categories, while the IAA features (b) exhibit noticeable confusion. The details are further explained **in supplement material A.3**.

To address the aforementioned challenges, we develop an enhancer against long-tail for aesthetics-oriented models (ELTA) that encompasses three key perspectives: ***First***, unlike traditional pixel-level augmentation methods, our ELTA implements feature-level augmentation based on the mixup (Zhang et al., 2017) to augment the minority. We propose an improved sampling strategy to assign different sampling weights to the minority and majority, enabling efficient generation of minority features even in the absence of well-defined categories. ***Second***, Gong et al. (Gong et al., 2022) state that a well-trained model should exhibit a key characteristic: items closer in label space should also be nearer in feature space. Inspired by this idea, a Feature-Label Similarity Alignment module is proposed to maximize the consistency between feature similarity and label similarity. Specifically, we take similarity as a metric for measuring distance, leverage label similarity to refine feature similarity, and optimize the model's representation learning. ***Third***, we introduce an Adaptive Probability Distribution Sharpening module to alleviate the smoothness of predicted distributions. Specifically, it is oriented by the adaptive magnitude of temperature scaling and the learnable pseudo-label selection threshold to obtain relatively accurate pseudo-labels. Our contributions are concluded as follows:

- The issue of long-tailed distribution in IAA datasets is revealed, highlighting its specificity and severe negative effects.
- Our proposed ELTA mitigates the data imbalance by augmenting minority features, aligning features to labels, and improving pseudo-labeling accuracy. To our knowledge, this is the first solution proposed against long-tail for aesthetics-oriented models.
- ELTA achieves state-of-the-art performance on four representative IAA datasets. Furthermore, the integration of ELTA with existing approaches can be seamlessly achieved, resulting in significant improvements in performance.

2. Related Works

Image Aesthetics Assessment and its long-tailed issue. General IAA encompasses three types of tasks: aesthetic binary classification (Datta et al., 2006; Luo & Tang, 2008), aesthetic score regression (Ma et al., 2017; Sheng et al., 2018; He et al., 2022; 2023a) and score distribution prediction (Talebi & Milanfar, 2018; Chen et al., 2020; She et al., 2021; Ke et al., 2021; Tu et al., 2022; He et al., 2023b), while personalized IAA adopts an aesthetic model for individual user's preference (Ren et al., 2017; Lv et al., 2021; Yang et al., 2022). Traditional methods rely on manually designed and extracted features from images. These visual features are then mapped to annotated labels via classifiers or regressors. In contrast, learning-based methods use Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs) with robust feature extraction capabilities to replace the manual process.

While existing learning-based IAA methods have achieved impressive results, they confront a critical issue: model bias caused by the long-tailed distribution of the datasets. This problem was first highlighted by He *et al.* (He et al., 2022) in their construction of the TAD66K dataset, where they employed a dedicated balancing strategy to mitigate it. However, a fundamental hurdle lies in the inherent nature of human annotation itself. The inclination of annotators to shy away from extreme ratings inevitably induces a long-tailed distribution, regardless of the dataset construction. Our work addresses this issue by proposing plug-and-play modules that leverage augmented minority features, consistent feature-label alignments, and accurate self-labeling to mitigate the long-tailed distribution.

Long-tailed Learning. Previous research on long-tailed learning has predominantly focused on classification tasks. ***Typical methods*** include re-sampling (Chawla et al., 2002; Kang et al., 2019; Wang et al., 2019; Xu et al., 2022; Bai et al., 2023) and re-weighting (Cui et al., 2019; Cao et al., 2019; Ren et al., 2020; Wang et al., 2021b; Du et al., 2023), which rebalance the contribution of each class. However, these methods are less applicable to aesthetics-oriented scoring tasks, which have no explicit class boundaries. ***Transfer learning methods*** (Yang & Xu, 2020; Liu et al., 2021; He et al., 2021; Chen & Su, 2023; Wei & Gan, 2023) aim to transfer knowledge from majority classes to enhance the learning of the minority. However, the highly abstract nature of aesthetics often hinders effective knowledge extraction and transfer. Some other methods use data augmentation techniques to expand datasets (Kim et al., 2020; Wang et al., 2021a; Ahn et al., 2023; Perrett et al., 2023), while they often risk unintentionally degrading aesthetic quality and introducing inconsistencies between the original labels and the augmented images.

Although a small number of recent researches have focused on ***long-tailed regression tasks*** (Yang et al., 2021; Gong et al., 2022), we find that these methods have not fully overcome the earlier mentioned challenges and their performance gains on IAA tasks are relatively limited. This limitation may stem from the methods' inadequate consideration of the IAA task's inherent complexities, which include subjectivity of aesthetic perception and the uncertainty of aesthetic labels. Instead, we fully consider these aesthetic properties and design a dedicated solution.

3. Method

3.1. Problem Setup

We consider two primary tasks, distribution prediction and score regression, based on whether the labels are in distributed or single-valued form. **1)** In the distribution prediction task, the training set with N samples is denoted as $\boldsymbol{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$. Here, $\boldsymbol{x}_i$ represents an input image, and the corresponding ground truth distribution of human ratings is indicated by the empirical probability mass function (PMF) $\boldsymbol{y}_i = \left[y_i^1, y_i^2, ..., y_i^C\right]$. Each element in $\boldsymbol{y}_i$ represents an aesthetic quality level, with C denoting the total number of these levels. The goal of model training is to predict the PMF $\hat{\boldsymbol{y}}_i = \left[\hat{y}_i^1, \hat{y}_i^2, ..., \hat{y}_i^C\right]$, which should be an accurate estimate of $\boldsymbol{y}_i$. For each image $\boldsymbol{x}_i$, the ground truth Mean Opinion Score (MOS) and the predicted score can be calculated as $s_i = \sum_{j=1}^{C} j \times y_i^j$ and $\hat{s}_i = \sum_{j=1}^{C} j \times \hat{y}_i^j$, respectively. **2)** In the score regression task, each image $\boldsymbol{x}_i$ is assigned a scalar score y_i. Despite this, the model can still output a distribution and then minimize the gap between the score $\hat{s}_i$ derived from the distribution and the ground truth label s_i.

IAA datasets inevitably exhibit long-tailed distributions, where a large number of images' aesthetic qualities cluster around the mean $\bar{s} = \frac{1}{N}\sum_{i=1}^{N} s_i$. This leads to a tendency for the models to output medium scores, which means the minority samples with labels that deviate substantially from the mean are prone to larger prediction errors, as indicated by $|\hat{s}_i - s_i| \propto |s_i - \bar{s}|$. Our primary focus will be on addressing these amplified errors observed in minority samples.

3.2. Tail Features Augmentation

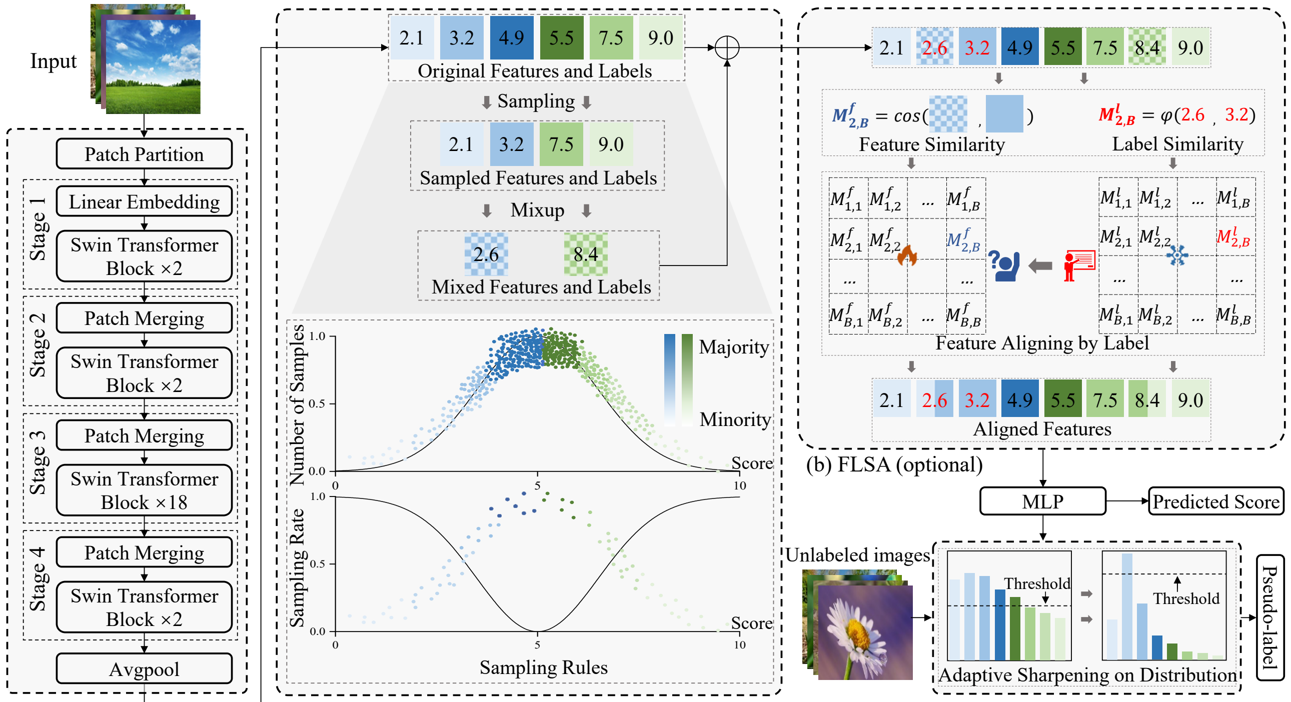
Image augmentation approaches such as cropping, flipping, and rotation, which directly manipulate the original image, can potentially destroy the intricate aesthetic information, creating an irreversible impact independent of the network's learning process (Hosu et al., 2019; Chen et al., 2020). Therefore,***feature-level enhancement is a more suitable augmentation approach for IAA***. Due to its operation on higher-level features extracted from images, rather than altering raw pixels, it preserves essential aesthetic qualities while generating new, diverse samples that better represent the minority.

As a classic technique, mixup (Zhang et al., 2017) can be applied to generate new samples and their associated labels through linear interpolation. Specifically, considering an image-label pair $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and $(\boldsymbol{x}_j, \boldsymbol{y}_j)$, let $\boldsymbol{z}_i^k, \boldsymbol{z}_j^k = f^k(\boldsymbol{x}_i), f^k(\boldsymbol{x}_j)$ be the features extracted at the k-th layer of a deep learning model from $\boldsymbol{x}_i, \boldsymbol{x}_j$. Then the mixed feature $\tilde{\boldsymbol{z}}^k$ and label $\tilde{\boldsymbol{y}}$ are generated as follows:

$$\begin{aligned} \tilde{\boldsymbol{z}}^k &= \lambda \boldsymbol{z}_i^k + (1-\lambda)\boldsymbol{z}_j^k, \\ \tilde{\boldsymbol{y}} &= \lambda \boldsymbol{y}_i + (1-\lambda)\boldsymbol{y}_j. \end{aligned} \tag{1}$$

It should be noted that when $k = 0$, the input feature $\boldsymbol{z}_i^0$ is identical to the original image $\boldsymbol{x}_i$, reverting to the initial version of mixup. In the following text, the superscript k will be omitted for brevity.

A mixup operation involves choosing two samples, and


Input
Patch Partition
Stage 1
Linear Embedding
Swin Transformer Block ×2
Stage 2
Patch Merging
Swin Transformer Block ×2
Stage 3
Patch Merging
Swin Transformer Block ×18
Stage 4
Patch Merging
Swin Transformer Block ×2
Avgpool
2.1 3.2 4.9 5.5 7.5 9.0
Original Features and Labels
Sampling
2.1 3.2 7.5 9.0
Sampled Features and Labels
Mixup
2.6 8.4
Mixed Features and Labels
Number of Samples
Majority
Minority
Score
Sampling Rate
Score
Sampling Rules
2.1 2.6 3.2 4.9 5.5 7.5 8.4 9.0
$M^f_{2,B} = cos($, $)$
Feature Similarity
$M^l_{2,B} = \varphi(2.6, 3.2)$
Label Similarity
$M^f_{1,1}$ $M^f_{1,2}$ ... $M^f_{1,B}$
$M^f_{2,1}$ $M^f_{2,2}$ $M^f_{2,B}$
$M^f_{B,1}$ $M^f_{B,2}$ ... $M^f_{B,B}$
$M^l_{1,1}$ $M^l_{1,2}$ ... $M^l_{1,B}$
$M^l_{2,1}$ $M^l_{2,2}$ $M^l_{2,B}$
$M^l_{B,1}$ $M^l_{B,2}$ ... $M^l_{B,B}$
Feature Aligning by Label
2.1 2.6 3.2 4.9 5.5 7.5 8.4 9.0
Aligned Features
MLP
Predicted Score
Unlabeled images
Threshold
Threshold
Adaptive Sharpening on Distribution
Pseudo-label

(a) TFA (optional)

(b) FLSA (optional)

(c) APDS (optional)

Figure 3: Architecture of the proposed ELTA model. ELTA consists of three **optional** components: (a) TFA performs mixup between instances sampled by our dedicated strategy to enhance minority features; (b) FLSA aligns features and labels based on their similarity consistency; (c) APDS reshapes the predictive distributions to work together with high thresholding for high-quality pseudo-label selection.

most of the previous work used random sampling. However, while random sampling may be effective in class-balanced datasets, it proves inadequate for long-tailed IAA datasets. This is due to the inherent dominance of majority instances within such datasets, resulting in an unintended abundance of majority features when employing random sampling and deviating from our original objective of enhancing minority representation.

To solve the above problem, we use the label information to guide more sampling of the minority and design a two-stage strategy. Initially, the sampling probability $P(i)$ for $\boldsymbol{z}_i$ is calculated based on its score deviation from the mean score of all samples. Samples far from the mean have high sampling probabilities, thus ensuring a bias towards the minority. Once the first sample feature $\boldsymbol{z}_i$ is chosen, the probability $P(j|i)$ that each of the remaining features is paired with the $\boldsymbol{z}_i$ can be calculated. This probability hinges on the score difference between the two samples. A smaller difference means a higher chance of pairing, thus ensuring the aesthetic similarity within pairs. The equations are detailed below:

$$P(i) = \frac{\exp\left(|\bar{s} - s_i|/\tau_1\right)}{\sum_{k=1}^{B} \exp\left(|\bar{s} - s_k|/\tau_1\right)}, \tag{2}$$

$$P(j|i) = \frac{\exp\left(\tau_2/|s_i - s_j|\right)}{\sum_{k=1,2,\cdots,i-1,i+1,\cdots,B} \exp\left(\tau_2/|s_i - s_k|\right)}, \tag{3}$$

where $\bar{s}$ represents the mean score of all the samples within that batch, and B is the batch size. τ_1 and τ_2 are temperature hyper-parameters that can be used to regulate the level of preference for the minority samples.

In traditional methods, the mixing factor λ in Eq. (1) is randomly selected from a beta distribution. To enhance the representation of minority data, We change λ to be obtained from the following equation:

$$\lambda = \frac{P(i)}{P(i) + P(j)} = \frac{\exp\left(|\bar{s} - s_i|/\tau_1\right)}{\exp\left(|\bar{s} - s_i|/\tau_1\right) + \exp\left(|\bar{s} - s_j|/\tau_1\right)}, \tag{4}$$

where $P(j)$ is computed by Eq. (2). This allocation of weights allows the sample which towards the minority more in the pair to have a dominant influence in the mixup process, reducing the impact of accidentally sampling a majority sample.

3.3. Feature-Label Similarity Alignment

In IAA tasks, features often exhibit significant confusion, as illustrated in Figure 2. This presents a challenge when addressing data imbalance through feature augmentation. Previous work (Gong et al., 2022) has shown items closer in label space should also be nearer in feature space. By leveraging this insight, label similarity can guide the refinement of features. This alignment process fosters a harmonious relationship between features and labels, promoting consistency and rectifying potential mismatches.

Specifically, the images $\boldsymbol{x}$ are first fed into the encoder, which extracts feature vectors $\boldsymbol{z}$ that capture their essential characteristics. These feature vectors are then compared pairwise using cosine similarity and yield a feature similarity matrix, denoted as $M^f \in \mathbb{R}^{B\times B}$, where the element of M^f can be formulated as:

$$M^f_{i,j} = \cos(\boldsymbol{z}_i, \boldsymbol{z}_j) = \frac{\boldsymbol{z}_i \cdot \boldsymbol{z}_j}{|\boldsymbol{z}_i| \cdot |\boldsymbol{z}_j|}. \tag{5}$$

Each element within this matrix holds a value that quantifies the similarity between a specific pair of feature vectors. Besides, the similarity matrix of labels $M^l \in \mathbb{R}^{B\times B}$ can be expressed as a function φ derived from their absolute differences:

$$M^l_{i,j} = \varphi(s_i, s_j) = 1 - \frac{|s_i - s_j|}{\max\limits_k s_k - \min\limits_k s_k}. \tag{6}$$

In particular, within the label similarity matrix, $M^l_{i,j} = 1$ signifies corresponding sample pairs that share identical labels, representing the highest level of similarity. Conversely, $M^l_{i,j} = 0$ indicates a complete absence of relevance. As features are adaptable through learning, the backpropagation process utilizes the label similarity matrix to guide the feature similarity matrix toward closer alignment, effectively minimizing the discrepancy between them. The alignment loss can be expressed as $\mathrm{MSE}(M^f, M^l)$. A tunable parameter, which controls the weight of this loss, is then multiplied with it. Subsequently, this product is added to the original supervised loss (e.g., EMD), acting as a regularization term.

3.4. Adaptive Probability Distribution Sharpening

While expanding minority from unlabeled samples offers a promising approach, how to properly utilize these samples requires careful consideration. We employ an Adaptive Probability Distribution Sharpening strategy to pseudo-label unlabeled data, specifically seeking minority samples to enrich diversity.

Challenges in selecting high-quality minority samples. Existing research (Berthelot et al., 2019; Sohn et al., 2020; Zhang et al., 2021) demonstrates that setting a high threshold and retaining only predictions with confidence exceeding this threshold can effectively filter out samples with large prediction errors. However, this approach confronts challenges due to the subjective and continuous nature of IAA labels, which leads to a softer, smoother, and less distinct predicted distribution. Consequently, identifying an optimal threshold for selecting high-quality minority samples is difficult. Especially, setting the threshold too high may exclude almost all potential samples, whereas a too-low threshold could include numerous samples with significant errors.

Reshaping the distribution adaptively. To overcome the challenges posed by IAA's smooth predicted distribution, we propose a temperature scaling technique inspired by (Hinton et al., 2015). By adjusting the original distribution, we aim to facilitate effective threshold selection for identifying reliable pseudo-labels. Specifically, given the logit vector $\boldsymbol{z}_i$, the new prediction confidence distribution $\hat{\boldsymbol{y}}_i$ is obtained as follows:

$$\hat{\boldsymbol{y}}_i = \mathrm{Softmax}(\boldsymbol{z}_i/\tau) = \frac{e^{\boldsymbol{z}_i/\tau}}{\sum_{j=1}^{B} e^{\boldsymbol{z}_j/\tau}}, \tag{7}$$

where τ is a temperature parameter that scales logit vectors. Increasing τ above 1 smooths the probability distribution, while below 1 sharpens the distribution.

To prioritize reshaping minority logits and ensure they can exceed the threshold more easily, we introduce an adaptive temperature that adjusts the smoothness based on the individual score of each sample. This adaptivity is achieved by modifying the temperature factor τ in Eq. (7) to function $\tau_i(\beta)$ on the sample's score, and then the score-based temperature distribution can be expressed as:

$$\begin{aligned} \tau_i(\beta) &= e^{-\beta|\hat{s}_i - \bar{s}|}, \\ \hat{\boldsymbol{y}}_i &= \frac{e^{\boldsymbol{z}_i/\tau_i(\beta)}}{\sum_{j=1}^{B} e^{\boldsymbol{z}_j/\tau_j(\beta)}}, \end{aligned} \tag{8}$$

Here, $|\hat{s}_i - \bar{s}|$ measures the difference between a sample's predicted score $\hat{\boldsymbol{s}}_i$ and the average score. The hyper-parameter $\beta > 0$ controls the steepness of temperature decline as the score difference increases. This exponentially decreasing temperature function prioritizes sharpening for minority samples. As the score difference grows, the temperature approaches 0, resulting in a sharper distribution of the minority. Conversely, majority samples with smaller differences maintain a smoother distribution.

Determining hyper-parameter by grid search. Unlike previous works where the threshold is typically an independent hyper-parameter to derive pseudo-labels, we integrate the threshold and sharpening magnitude as interconnected hyper-parameters. To achieve this, we introduce an additional stage to explore diverse combinations of these two hyper-parameters via grid search. This grid search is guided by MAE between generated and ground-truth labels. The search goal is to pinpoint the specific hyper-parameters combination that yields the most reliable and accurate pseudo-labels.

Initially, we construct a small labeled dataset, which can be sampled randomly from the validation set. We then employ the trained model to assess the samples from this dataset and select pseudo-labels based on the APDS strategy. Following this, a set of pseudo-labeled samples is generated, represented as $D' = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) | \max(\hat{\boldsymbol{y}}_i(\beta)) \geq t\}_{i=1}^{M}$, where $\max(\hat{\boldsymbol{y}}_i(\beta)) \geq t$ denotes the condition for pseudo label screening. Next, we calculate the labeled score s_i and predicted score $\hat{s}_i$ for each sample in $D^{'}$ (noting that this is based on the formula $s_i = \sum_{j=1}^{C} j \times y_i^j$ and $\hat{s}_i = \sum_{j=1}^{C} j \times \hat{y}_i^j$). By adjusting the magnitude β and the selection threshold t , different pseudo-labels can be generated. Finally, we aim to find the parameters β^* and t^* that minimize the MAE through a grid search method, which can be expressed as:

$$\beta^*, t^* = \arg\min_{\beta,t} \sum_{i=1}^{M} |\hat{s}_i(\beta, t) - s_i| \tag{9}$$

Once this optimal configuration is identified, we leverage it to rigorously filter pseudo-labels for subsequent use in the self-training process.

3.5. Overall Architecture and Training Loss

Overall architecture. The architecture is shown in Figure 3. Our method aims to achieve generic enhancement, independent of the backbone choice. Therefore, we select the Swin Transformer V2 (Liu et al., 2022), known for its versatility and popularity, as our network backbone. In addition, We designed the modules with flexibility in mind. Each one is optional and can function independently. Their loose coupling allows for easy integration or removal as needed.

Training loss. Following the selection of suitable pseudo-labeled samples using the aforementioned strategy, these samples are then incorporated into the original dataset for subsequent training epochs. During the self-training process, the loss is categorized into two types: labeled loss L_{s}, derived from the original training set, and unlabeled loss L_{u}, sourced from the pseudo-labeled samples:

$$L_{\text{s}} = \frac{1}{B^l} \sum_{i=1}^{B^l} H\left(\boldsymbol{y}_i^l, f\left(\boldsymbol{x}_i^l\right)\right),$$
$$L_{\text{u}} = \frac{1}{B^u} \sum_{i=1}^{B^u} \mathbb{1}\left[\max\left(\hat{\boldsymbol{y}}_i^u(\beta^*)\right) \geq t^*\right] \cdot H\left(\hat{\boldsymbol{y}}_i^u, f\left(\boldsymbol{x}_i^u\right)\right), \tag{10}$$

where superscripts l and u denote labeled and unlabeled data, respectively. H denotes the loss function and $\hat{\boldsymbol{y}}_i^u$ represents the generated pseudo-labels. The indicator function, denoted by $\mathbb{1}[\cdot]$, assigns a value of 1 only when the maximum value of $\hat{\boldsymbol{y}}_i^u$ exceeds the threshold t^*; otherwise, its value is 0.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate the performance of our approach on four representative datasets: AVA (Murray et al., 2012), AADB (Kong et al., 2016), TAD66K (He et al., 2022), PARA (Yang et al., 2022), which are the general, multi-attribute, theme-oriented and personalized aesthetic datasets, respectively, for IAA tasks. More details about datasets and train-test split ways are shown **in supplement material A.1**.

Evaluation metrics. We adopt two well-known evaluation metrics, the pearson linear correlation coefficient (PLCC, $\mathcal{P}$) and the spearman rank correlation coefficient (SRCC, $\mathcal{S}$) to evaluate performance. While models trained on long-tailed datasets often struggle with minority samples, these samples typically make up a small portion of the datasets and thus have a limited reflection on both holistic SRCC and PLCC metrics. To analyze further, we sort each dataset by its ground-truth scores and split it into three categories based on percentiles: top 20% as "high", bottom 20% as "low", and the remaining 60% as "medium". We consider the "high" and "low" categories as minority samples and the "medium" category as the majority. Then the Mean Absolute Error (MAE) for low, medium, high categories are calculated, denoted as $\mathcal{L}$, $\mathcal{M}$, $\mathcal{H}$, respectively.

Benchmark models. We conduct a comparative analysis involving 7 SOTA IAA models: NIMA (Talebi & Milanfar, 2018), HGCN (She et al., 2021), BIAA (Zhu et al., 2020), TANet (He et al., 2022), MaxViT (Tu et al., 2022), MUSIQ (Ke et al., 2021), EAT (He et al., 2023b), and 2 Deep Imbalanced Regression (DIR) models (Yang et al., 2021; Gong et al., 2022), which excel at long-tailed regression tasks. To ensure fairness, we use the same Swin Transformer V2 (Liu et al., 2022) backbone for all DIR methods and ELTA. Since some models have not been tested on all four datasets, we retrained them using their publicly available code and recommended parameter settings. Training process is provided in **supplement material A.2**.

4.2. Performance Evaluations

4.2.1. Comparison with IAA methods

Table 1 presents performance comparisons with IAA methods. Three key observations emerge from these results: **1)** Our method demonstrates superior performance on the AVA, AADB, and PARA datasets compared to previous methods in terms of PLCC and SRCC correlation metrics. However, it slightly lags behind EAT on the TAD66K dataset because ELTA lacks theme awareness. Despite this shortcoming, it still ranks high against other methods. **2)** ELTA achieves SOTA performance in both $\mathcal{L}$ and $\mathcal{H}$. This suggests ELTA effectively reduces evaluation errors for the minority. **3)**

Table 1: Comparing our ELTA with 7 IAA methods on the four datasets. The 'Base.' represents a Swin Transformer baseline model removing all the modules proposed. The best result for each metric is bolded in **red**.

Dataset	Metric	CNN-based models				Transformer-based models				
		NIMA	HGCN	BIAA	TANet	MaxViT	MUSIQ	EAT	Base.	Ours
AVA	$\mathcal{P}\uparrow$	0.636	0.687	0.668	0.765	0.745	0.738	0.770	0.743	**0.777**
	$\mathcal{S}\uparrow$	0.612	0.665	0.651	0.758	0.708	0.726	0.759	0.735	**0.764**
	$\mathcal{L}\downarrow$	0.655	0.675	0.653	0.630	0.600	0.647	0.490	0.616	**0.438**
	$\mathcal{M}\downarrow$	0.322	0.321	0.382	**0.237**	0.317	0.305	0.313	0.295	0.302
	$\mathcal{H}\downarrow$	0.648	0.660	0.568	0.729	0.531	0.628	0.433	0.513	**0.426**
AADB	$\mathcal{P}\uparrow$	0.711	0.734	0.733	0.742	0.748	0.761	0.767	0.740	**0.772**
	$\mathcal{S}\uparrow$	0.700	0.716	0.710	0.749	0.742	0.751	0.759	0.732	**0.760**
	$\mathcal{L}\downarrow$	1.450	1.453	1.508	1.394	1.592	1.447	1.375	1.526	**1.289**
	$\mathcal{M}\downarrow$	0.874	0.989	0.897	0.846	**0.782**	0.880	0.828	0.896	0.905
	$\mathcal{H}\downarrow$	1.527	1.299	1.423	1.355	1.461	1.159	1.260	1.402	**1.141**
TAD66K	$\mathcal{P}\uparrow$	0.405	0.493	0.431	0.531	0.513	0.517	**0.546**	0.507	0.539
	$\mathcal{S}\uparrow$	0.390	0.486	0.417	0.513	0.484	0.489	**0.517**	0.478	0.496
	$\mathcal{L}\downarrow$	1.851	1.808	1.734	1.598	1.570	1.627	1.591	1.621	**1.457**
	$\mathcal{M}\downarrow$	0.812	0.780	0.876	**0.682**	0.746	0.728	0.782	0.793	0.812
	$\mathcal{H}\downarrow$	1.690	1.370	1.669	1.651	1.402	1.460	1.175	1.354	**1.162**
PARA	$\mathcal{P}\uparrow$	0.862	0.881	0.886	0.899	0.936	0.918	0.940	0.925	**0.943**
	$\mathcal{S}\uparrow$	0.877	0.865	0.858	0.887	0.902	0.899	0.909	0.897	**0.912**
	$\mathcal{L}\downarrow$	0.616	0.573	0.469	0.551	0.383	0.572	0.336	0.402	**0.327**
	$\mathcal{M}\downarrow$	0.344	0.290	0.328	0.299	0.282	0.315	0.276	0.314	**0.251**
	$\mathcal{H}\downarrow$	0.486	0.502	0.503	0.429	0.276	0.424	**0.256**	0.379	0.290

Our method slightly compromises some majority performance, which is a common challenge in other long-tail solutions(Zhou et al., 2023).

4.2.2. Comparison with DIR methods

The results of the comparison with DIR methods are shown in Table 2. Our approach outperforms the methods presented in (Yang et al., 2021) and (Gong et al., 2022) in two aspects. First, it achieves SOTA in the overall performance, PLCC and SRCC. Second, it excels in prediction for minority samples, indicating a significant improvement in this area.

4.3. Enhancing other Methods by Plugging

To demonstrate the effectiveness of our proposed ELTA and its seamless integration into existing methods, ELTA is plugged into the other seven IAA methods. As illustrated in Figure 4, we integrate FLSA and TFA as an overall optimization module for model representation learning; following this, we incorporate self-training with APDS module to enhance the final performance. Results are detailed in Table 3. The data indicates improvements in the PLCC, SRCC, and MAE in both low and high segments.

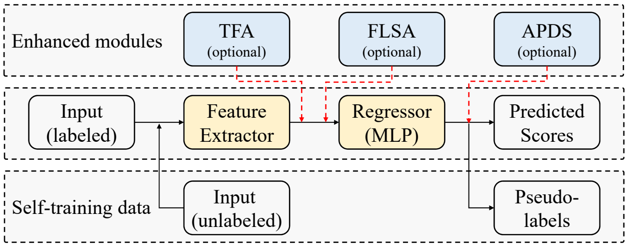


Figure 4: Example of plugging our modules into other methods.

4.4. Ablations and Analysis

Effectiveness of the modules. To assess the effectiveness of the modules, we create eight combinations by selecting whether to use each of the three modules. The results on the AVA dataset are shown in Table 4. We observe that the TFA module has a relatively significant effect on reducing minority errors, while FLSA and APDS demonstrate balanced improvements across all metrics. For example, removing TFA module (row 7) leads to an increase in MAE by 24.9% for the low segment and 8.7% for the high segment, in contrast to the configuration with all three modules active (row 8).

Table 2: Comparing our ELTA with 2 DIR methods, L&FDS (Yang et al., 2021) and Ranksim (Gong et al., 2022).

Dataset	Metric	Model			
		L&FDS	RankSim	Base.	Ours
AVA	$\mathcal{P}\uparrow$	0.752	0.759	0.743	**0.777**
	$\mathcal{S}\uparrow$	0.740	0.753	0.735	**0.764**
	$\mathcal{L}\downarrow$	0.588	0.542	0.616	**0.438**
	$\mathcal{M}\downarrow$	0.303	0.297	**0.295**	0.302
	$\mathcal{H}\downarrow$	0.509	0.485	0.513	**0.426**
AADB	$\mathcal{P}\uparrow$	0.746	0.750	0.740	**0.772**
	$\mathcal{S}\uparrow$	0.735	0.738	0.732	**0.760**
	$\mathcal{L}\downarrow$	1.473	1.537	1.526	**1.289**
	$\mathcal{M}\downarrow$	0.888	**0.881**	0.896	0.905
	$\mathcal{H}\downarrow$	1.407	1.295	1.402	**1.141**
TAD66K	$\mathcal{P}\uparrow$	0.509	0.514	0.507	**0.539**
	$\mathcal{S}\uparrow$	0.483	0.487	0.478	**0.496**
	$\mathcal{L}\downarrow$	1.611	1.560	1.621	**1.457**
	$\mathcal{M}\downarrow$	**0.787**	0.796	0.793	0.812
	$\mathcal{H}\downarrow$	1.369	1.343	1.354	**1.162**
PARA	$\mathcal{P}\uparrow$	0.933	0.930	0.925	**0.943**
	$\mathcal{S}\uparrow$	0.904	0.906	0.897	**0.912**
	$\mathcal{L}\downarrow$	0.371	0.363	0.402	**0.327**
	$\mathcal{M}\downarrow$	0.301	0.322	0.314	**0.251**
	$\mathcal{H}\downarrow$	0.360	0.345	0.379	**0.290**

Table 3: Cross-architecture evaluations are conducted to enhance other IAA methods, resulting in improved results on **AVA**.

Model	+ Module			Metric				
	TFA	FLSA	APDS	$\mathcal{P}\uparrow$	$\mathcal{S}\uparrow$	$\mathcal{L}\downarrow$	$\mathcal{M}\downarrow$	$\mathcal{H}\downarrow$
NIMA				0.636	0.612	0.655	**0.322**	0.648
	✓	✓		0.649	0.631	0.614	0.338	0.628
	✓	✓	✓	**0.658**	**0.640**	**0.593**	0.340	**0.601**
HGCN				0.687	0.665	0.675	**0.321**	0.660
	✓	✓		0.700	0.675	0.582	0.343	0.593
	✓	✓	✓	**0.714**	**0.693**	**0.564**	0.329	**0.575**
BIAA				0.668	0.651	0.653	0.382	0.568
	✓	✓		0.682	0.675	0.596	0.390	0.514
	✓	✓	✓	**0.699**	**0.687**	**0.544**	**0.381**	**0.497**
MUSIQ				0.738	0.726	0.647	**0.305**	0.628
	✓	✓		0.748	0.734	0.603	0.311	0.562
	✓	✓	✓	**0.761**	**0.745**	**0.588**	0.327	**0.482**
MaxViT				0.745	0.708	0.600	**0.317**	0.531
	✓	✓		0.750	0.728	0.557	0.340	**0.426**
	✓	✓	✓	**0.759**	**0.742**	**0.532**	0.333	0.428
TANet				0.765	0.758	0.630	**0.237**	0.729
	✓	✓		0.772	0.767	0.591	0.244	0.633
	✓	✓	✓	**0.779**	**0.771**	**0.562**	0.251	**0.585**
EAT				0.770	0.759	0.490	0.313	0.433
	✓	✓		0.777	0.765	**0.450**	0.310	0.426
	✓	✓	✓	**0.780**	**0.768**	**0.450**	**0.307**	**0.413**

Table 4: Ablation of different modules on the **AVA** dataset.

Module			Metric				
TFA	FLSA	APDS	$\mathcal{P}\uparrow$	$\mathcal{S}\uparrow$	$\mathcal{L}\downarrow$	$\mathcal{M}\downarrow$	$\mathcal{H}\downarrow$
			0.743	0.735	0.616	0.295	0.513
✓			0.749	0.740	0.518	0.323	0.474
	✓		0.758	0.752	0.592	0.287	0.520
		✓	0.760	0.749	0.575	0.286	0.480
✓	✓		0.764	0.748	0.484	0.324	0.431
✓		✓	0.762	0.756	0.461	0.314	0.445
	✓	✓	0.771	0.757	0.547	**0.286**	0.463
✓	✓	✓	**0.777**	**0.764**	**0.438**	0.302	**0.426**

Ablations on sampling strategy in TFA. Two ablation tests are carried out to validate the necessity of the proposed sampling strategy in the TFA module. The first employs a random sampling strategy. The results indicated a 4.6% decrease in both PLCC and SRCC. Through a look at the labels associated with the mixed features, we find that the augmentation occurs mainly in the majority rather than in the minority. In the second test, we set thresholds to distinguish between majority and minority, allowing only minority features defined in this way to be mixed. Despite experimenting with various thresholds, the optimal results achieved were 0.755 and 0.747, still below the 0.777 and 0.764 garnered with our proposed strategy.

Ablations on hyper-parameters. In Eq. (2) and (3), two temperature hyper-parameters are used to adjust the sampling probability distribution. They determine how much we favor the minority when selecting samples. To quantitatively assess how these hyper-parameters affect results, we conduct tests with various hyper-parameter values. The outcomes on AVA are shown in Figure 5. Notably, the SRCC metric reaches its peak with τ_1 set at 0.2 and τ_2 at 4.0.

5. Conclusion

This paper reveals the long-tailed distribution in the IAA datasets, its specificity, and the consequent model bias. To address this issue, we propose ELTA, which comprises three components: enhancing the minority features, aligning features and labels based on similarity consistency, and improving model self-training through an optimized pseudo-labeling strategy. ELTA outperforms existing IAA and DIR methods on four representative datasets, particularly in reducing the prediction errors of minority samples. However, like other solutions, ELTA encounters a minor decline in majority performance. In future work, we will continue to explore comprehensive performance enhancements without compromising the majority.

Acknowledgments

This work was supported by the Funds for the Innovation Research Group Project of NSFC (61921003).

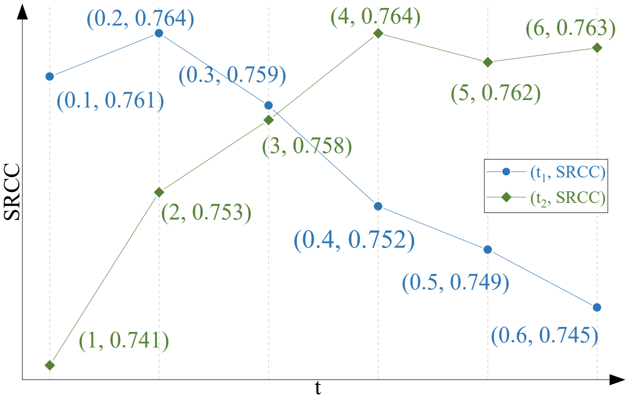


Figure 5: Effect of temperature hyper-parameters (τ_1 and τ_2) on SRCC.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ahn, S., Ko, J., and Yun, S.-Y. Cuda: Curriculum of data augmentation for long-tailed recognition. *arXiv preprint arXiv:2302.05499*, 2023.

Bai, J., Liu, Z., Wang, H., Hao, J., Feng, Y., Chu, H., and Hu, H. On the effectiveness of out-of-distribution data in self-supervised long-tail learning. *arXiv preprint arXiv:2306.04934*, 2023.

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., and Raffel, C. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Chen, J. and Su, B. Transfer knowledge from head to tail: Uncertainty calibration under long-tailed distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19978–19987, 2023.

Chen, Q., Zhang, W., Zhou, N., Lei, P., Xu, Y., Zheng, Y., and Fan, J. Adaptive fractional dilated convolution network for image aesthetics assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14114–14123, 2020.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. Studying aesthetics in photographic images using a computational approach. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part III 9*, pp. 288–301. Springer, 2006.

Du, F., Yang, P., Jia, Q., Nan, F., Chen, X., and Yang, Y. Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15814–15823, 2023.

Gong, Y., Mori, G., and Tung, F. RankSim: Ranking similarity regularization for deep imbalanced regression. In *International Conference on Machine Learning (ICML)*, 2022.

He, S., Zhang, Y., Xie, R., Jiang, D., and Ming, A. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 942–948, 2022.

He, S., Ming, A., Li, Y., Sun, J., Zheng, S., and Ma, H. Thinking image color aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21838–21847, 2023a.

He, S., Ming, A., Zheng, S., Zhong, H., and Ma, H. Eat: An enhancer for aesthetics-oriented transformers. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1023–1032, 2023b.

He, Y.-Y., Wu, J., and Wei, X.-S. Distilling virtual examples for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 235–244, 2021.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Hosu, V., Goldlucke, B., and Saupe, D. Effective aesthetics prediction with multi-level spatially pooled features. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9375–9383, 2019.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Ke, J., Wang, Q., Wang, Y., Milanfar, P., and Yang, F. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Kim, J., Jeong, J., and Shin, J. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13896–13905, 2020.

Kong, S., Shen, X., Lin, Z., Mech, R., and Fowlkes, C. Photo aesthetics ranking network with attributes and content adaptation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 662–679. Springer, 2016.

Liu, B., Li, H., Kang, H., Hua, G., and Vasconcelos, N. Gistnet: a geometric structure transfer network for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8209–8218, 2021.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12009–12019, 2022.

Luo, Y. and Tang, X. Photo and video quality evaluation: Focusing on the subject. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III 10*, pp. 386–399. Springer, 2008.

Lv, P., Fan, J., Nie, X., Dong, W., Jiang, X., Zhou, B., Xu, M., and Xu, C. User-guided personalized image aesthetic assessment based on deep reinforcement learning. *IEEE Transactions on Multimedia*, 2021.

Ma, S., Liu, J., and Wen Chen, C. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4535–4544, 2017.

Murray, N., Marchesotti, L., and Perronnin, F. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2408–2415. IEEE, 2012.

Park, S., Hong, Y., Heo, B., Yun, S., and Choi, J. Y. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6887–6896, 2022.

Perrett, T., Sinha, S., Burghardt, T., Mirmehdi, M., and Damen, D. Use your head: Improving long-tail video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2415–2425, 2023.

Ren, J., Shen, X., Lin, Z., Mech, R., and Foran, D. J. Personalized image aesthetics. In *Proceedings of the IEEE international conference on computer vision*, pp. 638–647, 2017.

Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33: 4175–4186, 2020.

She, D., Lai, Y.-K., Yi, G., and Xu, K. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8475–8484, 2021.

Sheng, K., Dong, W., Ma, C., Mei, X., Huang, F., and Hu, B.-G. Attention-based multi-patch aggregation for image aesthetic assessment. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 879–886, 2018.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Talebi, H. and Milanfar, P. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8): 3998–4011, 2018.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., and Li, Y. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pp. 459–479, 2022.

Wang, J., Lukasiewicz, T., Hu, X., Cai, J., and Xu, Z. Rsg: A simple but effective module for learning imbalanced datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3784–3793, 2021a.

Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C. C., and Lin, D. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, pp. 9695–9704, 2021b.

Wang, Y., Gan, W., Yang, J., Wu, W., and Yan, J. Dynamic curriculum learning for imbalanced data classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5017–5026, 2019.

Wei, C., Sohn, K., Mellina, C., Yuille, A., and Yang, F. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10857–10866, 2021.

Wei, T. and Gan, K. Towards realistic long-tailed semi-supervised learning: Consistency is all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3469–3478, 2023.

Xu, Y., Li, Y.-L., Li, J., and Lu, C. Constructing balance from imbalance for long-tailed image recognition. In *European Conference on Computer Vision*, pp. 38–56. Springer, 2022.

Yang, Y. and Xu, Z. Rethinking the value of labels for improving class-imbalanced learning. *Advances in neural information processing systems*, 33:19290–19301, 2020.

Yang, Y., Zha, K., Chen, Y., Wang, H., and Katabi, D. Delving into deep imbalanced regression. In *International Conference on Machine Learning*, pp. 11842–11851. PMLR, 2021.

Yang, Y., Xu, L., Li, L., Qie, N., Li, Y., Zhang, P., and Guo, Y. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19861–19869, 2022.

Yu, S., Guo, J., Zhang, R., Fan, Y., Wang, Z., and Cheng, X. A re-balancing strategy for class-imbalanced classification based on instance difficulty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 70–79, 2022.

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., and Shinozaki, T. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34: 18408–18419, 2021.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhou, Z., Li, L., Zhao, P., Heng, P.-A., and Gong, W. Class-conditional sharpness-aware minimization for deep long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3499–3509, 2023.

Zhu, H., Li, L., Wu, J., Zhao, S., Ding, G., and Shi, G. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *TCYB*, 2020.

Zhu, H., Zhou, Y., Li, L., Li, Y., and Guo, Y. Learning personalized image aesthetics from subjective and objective attributes. *IEEE Transactions on Multimedia*, 2021.

A. Appendix.

A.1. The Details of Datesets

The AVA dataset is the most extensive available aesthetic dataset, containing over 250,000 images, and each image is associated with a distribution of scores in a range of 1-10. Similar to (Murray et al., 2012; Talebi & Milanfar, 2018; Chen et al., 2020; He et al., 2022), we use 235,528 images for model training and 20,000 images for testing. AADB is an aesthetic attribute dataset containing about 10,000 images, each scored from 1 to 5. Following the previous work (Kong et al., 2016; Zhu et al., 2020; 2021), we use the standard split with 8,500 for training, 1,000 for testing and 500 for validation. The TAD66K dataset contains about 66,000 images covering 47 popular themes, and each image has been annotated with score from 1 to 10 on dedicated theme evaluation criteria. We use the official train-test split way (He et al., 2022; 2023b), 52,248 for training and 14,079 for testing. PARA is a personalized image aesthetics dataset with rich attributes, which consists of 31,220 images with annotated scores from 1 to 5. The train-test split way is the same with the work (Yang et al., 2022).

A.2. Training Details

Since the ground truth in the AVA dataset consists of the score distribution, we use the earth mover's distance (EMD) loss to measure the distance between the ground-truth and the predicted distribution. For the AADB, TAD66K, and PARA datasets, we use mean squared error (MSE) loss, as these datasets only provide holistic scores. The training process is optimized using the Adam optimizer, with a batch size of 48. Once the first round of training is completed, model evaluation is performed to select pseudo-labels. We use these pseudo-labels for only one additional round of self-training. When the model is trained on one labeled dataset, the unlabeled dataset used to provide the pseudo-labeled samples is randomly sampled and constructed from other datasets with no more instances than the original training dataset.

A.3. Details of Figure 2

In Figure 2, we present a comparative analysis of feature distributions learned by a classification model and an IAA model, using a toy experiment. To achieve this comparison, a small dataset with both category labels and aesthetic score labels is required. Considering that the theme labels in the TAD66K dataset can be used as category information, we sample the images from the TAD66K. The dataset obtained from this sampling adheres to a key characteristic: identical shapes in both category distribution and aesthetic score distribution. Specifically, we choose 10 categories out of the 47 available and divide aesthetic scores into ten equal intervals with a unit distance. In our constructed dataset, the sample size for the n-th category corresponds to the n-th score interval, which ensures consistency in the dataset distributions for the two different tasks. In addition, we employ identical experimental parameters and the same Swin Transformer V2 backbone network for feature extraction, differing only in the fully connected layers at the end of the network. After training the models, we perform dimensionality reduction on the extracted features and visualize them in Figure 2. It reveals clear boundaries between different categories in the classification task, while features in the IAA task appear more chaotic. By eliminating other interfering factors, including dataset distribution and backbone network, we confirm that this severe mismatch between features and labels is indeed attributable to the intrinsic characteristics of aesthetic tasks.

A.4. Mixup strategy

In Section 3.2, we point out that feature-level enhancement is a more suitable augmentation approach for IAA. Here we give some quantitative experimental results. We maintained the same sample selection strategy for mixup but shifted the mixup process from the feature level to the pixel level on the original images. The results in Table 5 indicate a noticeable decline in performance, whether we applied the TFA modules individually or all modules.

Table 5: Comparisons of feature-level augmentation and pixel-level augmentation on **AVA**.

Model	$\mathcal{P}\uparrow$	$\mathcal{S}\uparrow$	$\mathcal{L}\downarrow$	$\mathcal{M}\downarrow$	$\mathcal{H}\downarrow$
TFA(pixel)	0.731	0.719	0.597	0.368	0.541
TFA(feature)	0.749	0.740	0.518	0.323	0.474
TFA(pixel)+FLSA+APDS	0.752	0.741	0.487	0.340	0.485
TFA(feature)+FLSA+APDS	**0.777**	**0.764**	**0.438**	**0.302**	**0.426**

A.5. Image examples

Figure 6 presents some examples from both the AVA and TAD66K datasets. The Baseline model, marked in blue, struggles to effectively differentiate the aesthetic quality of images. Typically, its evaluation results cluster within a 4-6.5 point range, which means significant evaluation errors for images with high or low aesthetic values. In contrast, our proposed ELTA model, indicated in red, demonstrates improved performance. The result is closer to ground truth, which is represented in **black**.

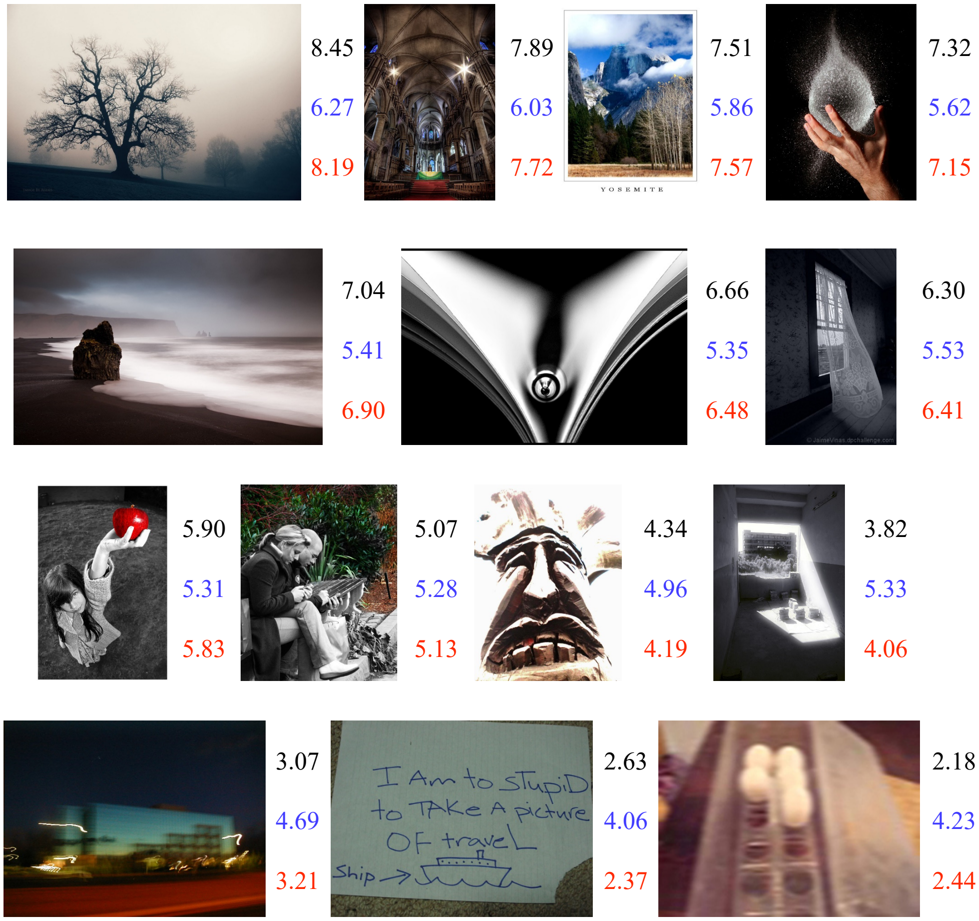


Figure 6: Images examples with their corresponding ground-truth and predicted scores.