KEEP KNOWLEDGE IN PERCEPTION: ZERO-SHOT IMAGE AESTHETIC ASSESSMENT

Guolong Wang$^{1*\dagger}$ Yike Tan$^{1*}$ Hangyu Lin$^1$ Chuchun Zhang$^1$

1 University of International Business and Economics

ABSTRACT

Image aesthetic assessment is an important issue in multimedia, but most existing studies employ supervised learning methods that rely on large-scale annotated data. However, aesthetic scoring annotations are difficult to obtain in large quantities. Therefore, this paper explores zero-shot image aesthetic assessment. We predict aesthetic scores by introducing knowledge of different attributes (e.g., Focus). First, we use prompt tuning to obtain a unique prompt for each aesthetic attribute as external knowledge. Second, we leverage image relations considering sentiment polarity as internal knowledge. Specifically, we obtain aesthetic attribute representations from pre-trained models via prompt learning, then select anchor images on specific attributes by sentiment polarity, computing aesthetic scores. Notably, annotated aesthetic scores are not used in the process. Experiments show that our zero-shot approach outperforms many comparisons using only a few anchor images.

Index Terms— Image aesthetic assessment, zero-shot learning, external knowledge, internal knowledge

1. INTRODUCTION

Image aesthetic assessment (IAA) is a significant task within the field of multimedia [1]. The feasibility of this task has been progressively augmented by advancements in deep learning methods based on large-scale databases [2, 3]. This progress is particularly notable given the widespread availability of aesthetic-related images and associated comments on the internet, which provides a convenient resource for executing this task [4]. With the help of these models, retailers or brands can estimate a score for their products automatically.

However, it is still challenging to estimate image aesthetics accurately due to the high-dimensional image features and the subjective nature of user behavior. Initially, image aesthetic assessment was defined as a supervised task based on large annotated corpora where images were labeled with a binary score or Likert scale. The existing methods always focus on fitting ground-truth scores by designing complex structures. Some backbone neural networks have achieved high performance. For instance, VGG network [5] has achieved 91.93% [6] accuracy on the CUHKPQ dataset [7].

Although these supervised methods have achieved significant performance, they are hindered by some limitations. Firstly, they lacked generalization ability as they depended on the fit of specific data, while generalization ability is crucial for a practical method [8]. The pattern of correlations between image features and aesthetic scores is rather heterogeneous across different datasets [9]. Secondly, the annotated single score in current databases proved insufficient to meet the evolving needs. These limitations drove the demand for zero-shot image aesthetic assessment, i.e., no ground-truth score is needed for training

Knowledge is crucial in a zero-shot setting. Semantic information is always used as knowledge, such as aesthetic attributes [11]. Some methods designed a multi-task framework by introducing a related task, such as emotion prediction [12]. Others designed network structures for a specific attribute [13]. Except for aesthetic attributes, some methods formulate external knowledge (e.g., object information [14] and image critiques [15]) and internal knowledge (e.g., self-supervision [16]).

In this work, we propose a Knowledge-enhanced Zero-shot Image Aesthetic Assessment (KZIAA) framework to address the limitations. Image aesthetic assessment is a subjective task where more semantic information than a score is useful [17, 18]. Consequently, attention shifted towards the finer-grained aspects of the aesthetic assessment of images [19, 20]. Thus, assessing different aspects of an image’s aesthetics became increasingly valuable, desiring knowledge introduction. It can offer comprehensive information and enhance the generalization ability of models [21]. This motivated our two techniques. Firstly, we improve the generalization ability of our method by utilizing prior knowledge from large models, in less it is over-fitted on a data distribution. Secondly, we select aesthetic representative images (aka anchor images) and employ image relations to enrich the aesthetic representation. Thus, image aesthetics combines external knowledge and internal relationships. The framework has three main novelties:

1) We select a few image anchors and use image relations to analyze image aesthetics.

2) We construct a continuous prompt for each aesthetic

$^*$Equal contribution. $^\dagger$Corresponding Author. This work is supported by the Open Projects Program of State Key Laboratory of Multimodal Artificial Intelligence Systems.

$^1$In this work, we adopt a setting same as [10], which may be different from the zero-shot setting in other tasks.
attribute with a pre-trained model and combine prompts of all attributes to rate new images.

3) We incorporate external knowledge and internal relationships to distinguish between good and bad images, thereby obtaining an image’s score in a zero-shot manner.

2. PROBLEM FORMULATION

In this paper, we focus on zero-shot image aesthetic assessment by designing a model $\mathcal{M}(\cdot)$. Given an image $I$, $\mathcal{M}$ can predict the aesthetic score $s$ for this image:

$$s = \mathcal{M}(I)$$

(1)

where $I \in \mathbb{R}^{w \times h}$, $w$ and $h$ are width and height of the image.

During the training of $\mathcal{M}(\cdot)$, it has no access to ground-truth score $s$. Instead, we use a collected large-scale vision-language corpus $D_t = \{(I_k, t_k)\}_{k=1}^N$ as external knowledge base and a quadruplet set $G$ as internal knowledge. $N_t$ is the number of multi-modal pairs. $q = (I, w, A_i, e) \in G$ consists of an image $I$, an aesthetic assessment word $w$, an attribute $A_i$, and a sentiment polarity score $e$. For attribute $A_i$, $i = 1, \cdots, N_a$ and $N_a$ is the number of attributes. Following [19], we have seven aesthetic attributes for images (i.e., $N_a = 7$).

3. IMAGE AESTHETIC ASSESSMENT DESIGN

Our research design comprises three components: (1) External knowledge of aesthetic attributes with continuous prompt, (2) Internal knowledge of image relation with polarity, and (3) Aesthetic score inference based on external and internal knowledge, as shown in Fig. 1.

3.1. External Knowledge with Continuous Prompt

In this section, we use image critique features to learn continuous prompts that can adapt large-scale vision-language pre-trained models to image aesthetic assessment, known as prompt tuning. The process is shown in Fig. 2.

An intuitive solution is using handcrafted templates, such as “{Good} photo.” and “{Bad} photo.” [22]. However, the handcrafted templates are always sub-optimal, hindering the generalization capability. Thus, we utilize prompt tokens $[v]_i, i \in [1, N_p]$ that will be optimized to formulate the continuous prompt templates $F(\text{Attribute})$ as Eqn. 2. We tune the embeddings of these tokens specifically to the image and critique data, and other parameters are kept frozen. Note that we have no access to the ground-truth image aesthetic scores.

$$F(\text{Attribute}) = [v]_1[v]_2 \cdots [v]_{N_p}[\text{Attribute}]$$

(2)

where $N_p$ is the number of continuous prompt tokens.

In this setting, we propose an attribute-specific prompt template where each attribute has a unique context. Thus, the prompt can extract features related to specific aesthetic attributes from the pre-trained model. Given the vision-language data corpus $D_t$ and $N_a$ attributes, we fine-tune the prompt. Specifically, we use $E_I$ and $E_T$ to denote image and text encoder, respectively. $E_I$ and $E_T$ share the same network structure, following Transformer Encoder in [23].

We then fine-tune the prompts by the similarity of image features and text features. The similarity between image $I$ and attribute $A_i$, $i \in [0, N_a - 1]$ is calculated by cosine metric $\langle \cdot, \cdot \rangle$ commonly used in the literature. The probability of predicting that $I$ is attribute $A_i$ is:

$$p(y = A_i | I) = \frac{\exp(\langle E_I(I) \cdot E_T(F(A_i)) \rangle / \tau)}{\sum_j \exp(\langle E_I(I) \cdot E_T(F(A_j)) \rangle / \tau)}$$

(3)

where $\tau$ is the temperature parameter.

We use the cross entropy loss for attribute classification:

$$\mathcal{L} = - \sum_{i} N_p \ A_i^g \ \log(p(y = A_i | I))$$

(4)

where $A_i^g$ is the ground-truth attribute label.

We get a unique prompt $F(A_i)$ for each attribute through prompt tuning. Different from [22], our prompt is attribute specific, not unified for all attributes, as a unified prompt may lead to over-fitting and sub-optimal in different attributes.
3.2. Internal Knowledge with Polarity

We use the relationship between images as internal knowledge. People perform better in pairwise comparison than absolute scoring in image aesthetic test [24]. Through comparison, they can shrink the estimation uncertainty due to the implicit knowledge of the relationship between two images.

Motivated by these observations, we select anchor images and adopt the comparison with them to build internal knowledge. As shown in Fig. 3, we summarize the way to obtain a quadruplet $q = (I, w, A_i, e) \in G$. For the word $w$, we extract it from the critique on attribute $A_i$ for image $I$. For the sentiment polarity score $e \in \{-1, 1\}$, we measure it on the same critique. Thus, the relationship between two images is built on aesthetic assessment and sentiment polarity regarding different aesthetic attributes.

Then, we select anchor images with positive and negative polarity, denoted as $I_p$ and $I_n$. Specifically, we select images whose absolute value of sentiment polarity score is larger than threshold $T_e$ from $G$. It makes the anchor images contain rich reference aesthetic knowledge in a contrasting perspective. Compared to the knowledge base in [25], our method integrates multi-modal knowledge information containing the relationship between images.

3.3. Aesthetic Score Inference

In this section, we make aesthetic score inferences based on the external knowledge of attribute $F(\text{Attribute})$ and internal knowledge in $I_p$ and $I_n$ from $G$. Firstly, we calculate the similarity between the test image and anchor images in $I_p$ and $I_n$ on each attribute, respectively. Thus, given a test image $I$, we can get a similarity vector $v_s \in \mathbb{R}^{2 \times N_a}$. Subsequently, we obtain the final score $s$ by weighted summation of scores under different attributes. The score $s_i$ under $A_i$, the weight $w_i$ for $s_i$, and the final score $s$ is shown in Eqs. 5, 6, and 7.

$$s_i = \frac{e^{v_s^i \cdot 0}}{e^{v_s^i \cdot 0} + e^{v_s^i \cdot T}}$$

$$w_i = \frac{< E_t(I) \cdot E_T(F(A_i)) >}{\sum_j < E_t(I) \cdot E_T(F(A_j)) >}$$

$$s = \frac{\sum_i^N_n s_i w_i}{N_a}$$

4. EXPERIMENT

4.1. Database

AVA [26]. It has over 250,000 images. We use the mean value of all ratings for an image as its ground-truth aesthetic score. We evaluate models on 19,929 test images in our zero-shot setting, following [10].

Multi-Modal Aesthetic (MMA) Database. We collect image aesthetic critiques for images in AVA to train attribute prompts and construct the quadruplet set: (1) We collect critiques on the original website following [15] and use critiques in [18] as a supplement. (2) We eliminate incomplete sentences (e.g., ‘This is a’). (3) We classify the critiques into seven aspects according to keywords summarized by professional photographers. We show MMA’s statistics in Table 1. Note that our MMA only contains images in AWA’s training set without ground-truth scores.

Table 1. Main statistics of MMA. ‘critiquespos’ and ‘critiquesneg’ are short for positive and negative critiques, respectively.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>#photos</th>
<th>#critiques</th>
<th>#critiquespos</th>
<th>#critiquesneg</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Impression</td>
<td>8,026</td>
<td>19,762</td>
<td>17,340</td>
<td>2,422</td>
</tr>
<tr>
<td>Composition &amp; perspective</td>
<td>10,499</td>
<td>33,850</td>
<td>27,691</td>
<td>6,159</td>
</tr>
<tr>
<td>Color &amp; Lighting</td>
<td>10,049</td>
<td>31,274</td>
<td>25,886</td>
<td>5,388</td>
</tr>
<tr>
<td>Subject of photo</td>
<td>9,543</td>
<td>27,294</td>
<td>22,485</td>
<td>4,809</td>
</tr>
<tr>
<td>Depth of field</td>
<td>8,116</td>
<td>20,023</td>
<td>17,470</td>
<td>2,553</td>
</tr>
<tr>
<td>Focus</td>
<td>8,758</td>
<td>22,613</td>
<td>19,255</td>
<td>3,358</td>
</tr>
<tr>
<td>Use of camera, exposure</td>
<td>8,259</td>
<td>20,696</td>
<td>17,723</td>
<td>2,973</td>
</tr>
<tr>
<td>Total</td>
<td>63,250</td>
<td>175,512</td>
<td>147,850</td>
<td>27,662</td>
</tr>
</tbody>
</table>

4.2. Evaluation Metrics

We adopt five evaluation metrics in the literature: RMSE, SRCC, PLCC, ACC, and AUC.

RMSE, SRCC, PLCC: RMSE measures the difference between predicted and actual observed values. SRCC and PLCC refer to Spearman’s rank and Pearson linear correlation coefficients between the predicted and ground-truth mean scores, respectively. These two metrics measure models’ ability to rank different images on the aesthetic score.

ACC, AUC: We can reduce IAA to a binary classification problem. The images whose score is above five are aesthetic and otherwise not aesthetic. We calculate the accuracy. AUC is Area Under the Receiver Operating Characteristic (ROC) curve. It is a performance metric commonly used to evaluate binary classification models.
Table 2. Comparisons on the AVA test dataset in RMSE, SRCC, PLCC, ACC, AUC. The best results are highlighted in bold. The second-best SRCC and PLCC are underlined.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Year</th>
<th>RMSE ↓</th>
<th>SRCC ↑</th>
<th>PLCC ↑</th>
<th>ACC ↑</th>
<th>AUC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLSP [31]</td>
<td>2019</td>
<td>0.851</td>
<td>0.733</td>
<td>0.715</td>
<td>0.173</td>
<td>0.181</td>
</tr>
<tr>
<td>AFDC+SPP [32]</td>
<td>2020</td>
<td>0.521</td>
<td>0.649</td>
<td>0.671</td>
<td>0.832</td>
<td>–</td>
</tr>
<tr>
<td>MUSIQ-single [33]</td>
<td>2021</td>
<td>0.497</td>
<td>0.719</td>
<td>0.731</td>
<td>0.814</td>
<td>–</td>
</tr>
<tr>
<td>TAVAR [13]</td>
<td>2023</td>
<td>–</td>
<td>0.758</td>
<td>0.765</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>TAVAR</td>
<td>2023</td>
<td>–</td>
<td>0.725</td>
<td>0.736</td>
<td>–</td>
<td>0.851</td>
</tr>
<tr>
<td>Zero-shot methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLIP-IQA</td>
<td>2023</td>
<td>1.334</td>
<td>0.173</td>
<td>0.181</td>
<td>0.712</td>
<td>0.595</td>
</tr>
<tr>
<td>VILA (pre-trained)</td>
<td>2023</td>
<td>–</td>
<td>0.657</td>
<td>0.663</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>miniGPT-4 (fine-tuned)</td>
<td>2023</td>
<td>2.688</td>
<td>0.080</td>
<td>0.078</td>
<td>0.638</td>
<td>0.536</td>
</tr>
<tr>
<td>KZIAA (fine-tuned)</td>
<td>2023</td>
<td>1.078</td>
<td>0.446</td>
<td>0.454</td>
<td>0.733</td>
<td>0.715</td>
</tr>
</tbody>
</table>

4.3. Implementation Details

We use CLIP [27] as the basic vision-language pre-trained model. Specifically, we use ViT-B/32 for the image encoder and Transformer for the text encoder. We insert context tokens before aesthetic attributes to fine-tune the continuous prompt. We set \( N_p = 16 \) and the size of \( [\text{Attribute}] \) to 60 in Eqn. 2. During fine-tuning, we freeze the parameters of image encoder \( E_I \) and text encoder \( E_t \). For training, we employed the Adam optimizer [28] with a learning rate of \( 1\times10^{-3} \). The BERT we used to calculate the sentiment polarity score is pre-trained following [29]. We applied Min-Max Normalization on \( s_i \) and \( w_i \) in Eqn. 7 to scale the data. All of our models are implemented by PyTorch and trained under the environment of Ubuntu 20.04. Note that we train our KZIAA without any annotated scores.

4.4. Results

Performance on AVA. The proposed KZIAA is a zero-shot method. We compare our method with several SOTA supervised methods and two zero-shot baselines: VILA [10] and CLIP-IQA [22]. The former is pre-trained, while the latter is not. We also fine-tune miniGPT-4 [30] on our MMA and use it as a competitor. Table 2 shows that our KZIAA outperforms CLIP-IQA and miniGPT-4, not VILA. A possible explanation is that our model is not pre-trained on the same set with VILA. Notably, it has the potential to catch up with supervised methods, as we only use a few anchor images.

Ablation study on \( T_s \). As selecting anchor images threshold \( T_s \) is a vital hyper-parameter, we perform an ablation study on \( T_s \). The results summarized in Table 3 show that our KZIAA is insensitive to the threshold. It indicates that our method is feasible for expanded scenarios.

Qualitative Analysis of prompts \( F(\text{Attribute}) \). We show an exemplar of continuous prompts in Fig. 4 by visualizing \( w_i \) (Eqn. 6). These weights contribute to the final score calculation and enhance the interpretability of the score from the aesthetic attribute level.

4.5. Conclusions

In this work, we propose a simple yet effective zero-shot strategy for image aesthetic assessment, a data-hungry subjective task. We estimate the aesthetic score by leveraging external knowledge and internal knowledge. Firstly, we obtain a unique context for each aesthetic attribute by prompt tuning. Subsequently, we construct a quadruplet set with image relationships and utilize sentiment polarity to select anchor images. Finally, we estimate the score considering the information of different attributes. Experiment results indicate the superiority of the proposed method to some zero-shot baselines and the potential to approach supervised methods.
References


