KEEP KNOWLEDGE IN PERCEPTION: ZERO-SHOT IMAGE AESTHETIC ASSESSMENT

(Guolong Wang*, Yike Tan*, Hangyu Lin, Chuchun Zhang)

Yike Tan
University of International Business and Economics
17 Apr, 2024

1. Background and Motivation

2. Methods

3. Experiments and Analysis

4. Conclusion and Future Work

Background

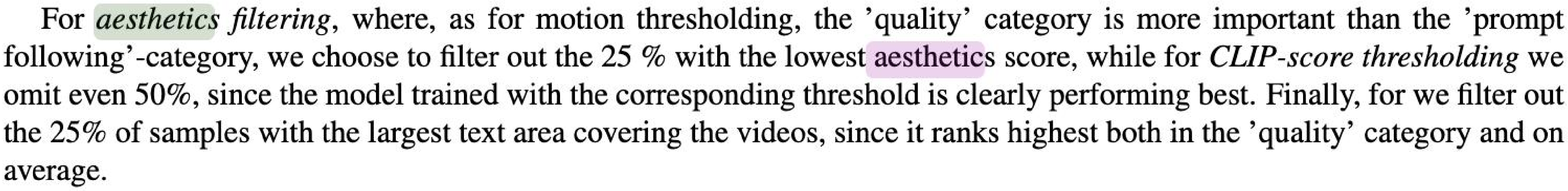
➢ Image Aesthetic Assessment (IAA)

This task aims to quantify the human perceived aesthetics of a given image.

➢ Applications:

Image recommendation, enhancement, retrieval, and generation [1] [2].

- Stable Video Diffusion



For *aesthetics filtering*, where, as for motion thresholding, the ’quality’ category is more important than the ’prompt following’-category, we choose to filter out the 25 % with the lowest aesthetics score, while for *CLIP-score thresholding* we omit even 50%, since the model trained with the corresponding threshold is clearly performing best. Finally, for we filter out the 25% of samples with the largest text area covering the videos, since it ranks highest both in the ’quality’ category and on average.

➢ Challenges:

- high-dimensional image features
- subjective nature

[1] Jiajing Zhang, Yongwei Miao, and Jinhui Yu, “A comprehensive survey on computational aesthetic evaluation of visual art images: Metrics and challenges,” IEEE Ac_x0002_cess, vol. 9, pp. 77164–77187, 2021.

[2] Blattmann, Andreas, et al. "Stable video diffusion: Scaling latent video diffusion models to large datasets." arXiv preprint arXiv:2311.15127 (2023).

Background

- ➢ Existing IAA methods, which primarily rely on human_x0002_labeled rating scores.
- ➢ Corressponding Limitations:
  - ▪ Lack generalization ability (depended on the fit of specific data)
  - ▪ Annotated data insufficient

Zero-Shot IAA

No ground-truth score is needed for training

Ground-Truth Score
6.93

Supervised

anchor images
& sentence

Zero-Shot

[1] *Nam et al., Zero-shot natural language video localization. ICCV 2021*

Motivation

Zero-Shot IAA

Two key ideas:

- **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., selfsupervision [16]).

- **Attention shifted towards the finergrained aspects**

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

[11~16] can be found in our reference part.

1. Background and Motivation

2. Methods

3. Experiments and Analysis

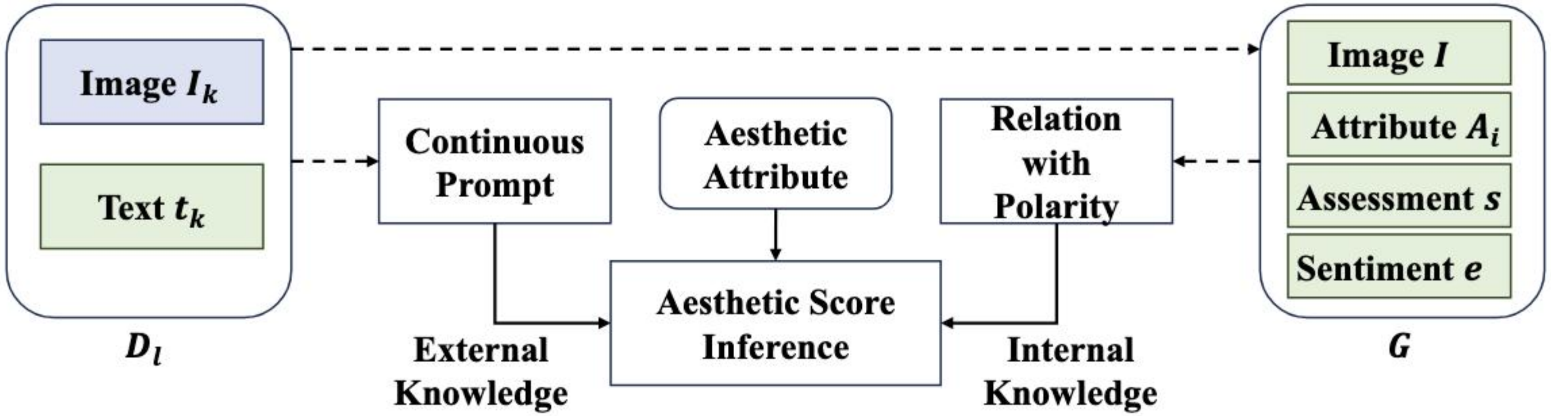
4. Conclusion and Future Work

Methods - Overall Framework

KZIAA (Knowledge-enhanced Zeroshot Image Aesthetic Assessment)

Three components:

- continuous prompt (external)
- relation with polarity (internal)
- aesthetic score inference.



Three main novelties:

- select a few **image anchors** and use **image relations** to enrich the aesthetic representation
- construct a **continuous prompt** for each aesthetic attribute (utilizing prior knowledge)
- incorporate **external knowledge and internal relationships** to distinguish between good and bad images, (obtaining an image's score in a zero-shot manner).

Methods - Multi-Modal Aesthetic Dataset (MMA)

- ➢ Download image comments of the AVA dataset from dpchallenge.com
- ➢ Design a set of aesthetic attribute keyword system to screen comments

General Impression	Subject of Photo	Composition &Perspective	Use of Camera, Exposure& Speed	Depth of Field	Color & Lighting	Focus
shot	theme	border	balance	length	coloring	softness
impression	rhythm	rotation	iso	blurriness	tune	attention
……	……	……	……	……		……
emotion	style	angle	lens	fuzz	histogram	sharpeness

- ➢ Filter and categorize comments based on keyword system

fig

Very simple, effective shot. → General Impression

~~**yeah baby!**~~

……

The colors are just great and the comp exudes such sensuality. Very good work. → Color & Lighting

Methods - External Knowledge with Continuous Prompt

➢ Prompt Tuning

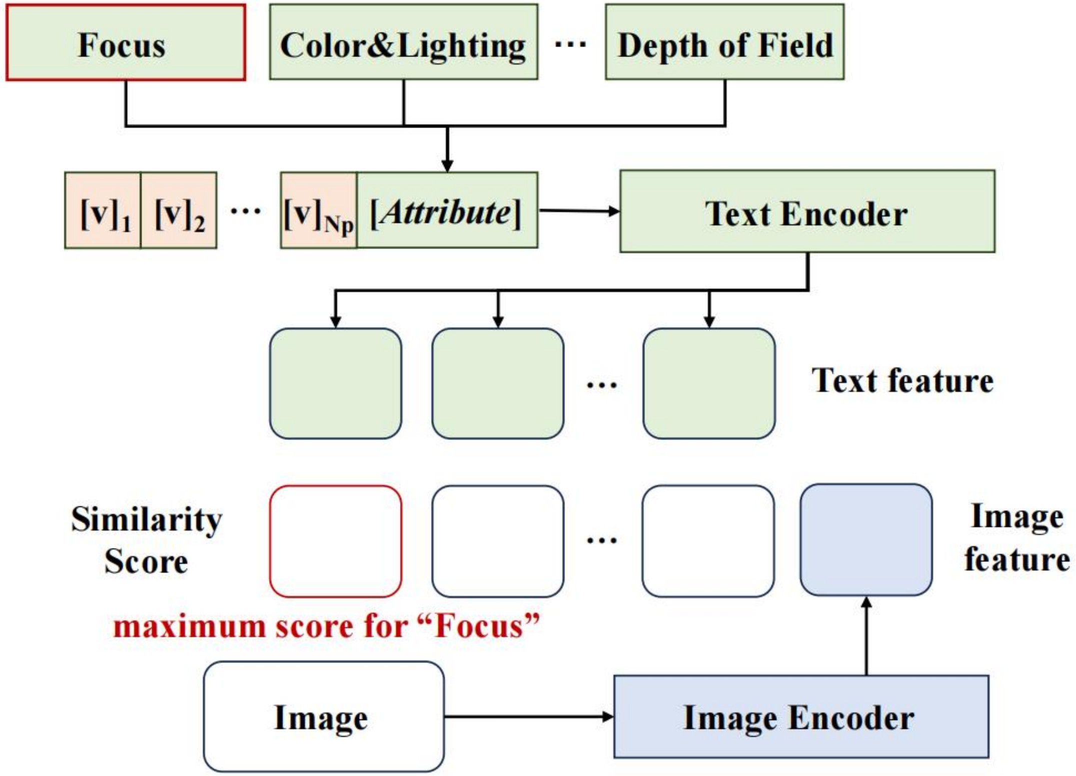
Design Prompts for each of the seven attributes.

$$\boldsymbol{F}(Attribute) = [\boldsymbol{v}]_1[\boldsymbol{v}]_2 \cdots [\boldsymbol{v}]_{N_p}[Attribute] \qquad (2)$$

Fine-tune the prompts by the similarity of image features and text features, obtain a unique context for each attribute.

$$p(y = A_i|I) = \frac{\exp(< E_I(I) \cdot E_T(\boldsymbol{F}(A_i)) > /\tau)}{\sum_j \exp(< E_I(I) \cdot E_T(\boldsymbol{F}(A_j)) > /\tau)}) \qquad (3)$$

$$\mathcal{L} = -\sum_i^{N_a} A_i^g \log(p(y = A_i|I)) \qquad (4)$$



➢ Weight Inference

Calculate the similarity between image features and the Prompt features of each attribute as the weight score

➢

$$w_i = \frac{< E_I(I) \cdot E_T(\boldsymbol{F}(A_i)) >}{\sum_j < E_I(I) \cdot E_T(\boldsymbol{F}(A_j)) >} \qquad (6)$$

Methods - Internal Knowledge with Polarity

- 1. Generate a quadruplet for each comment

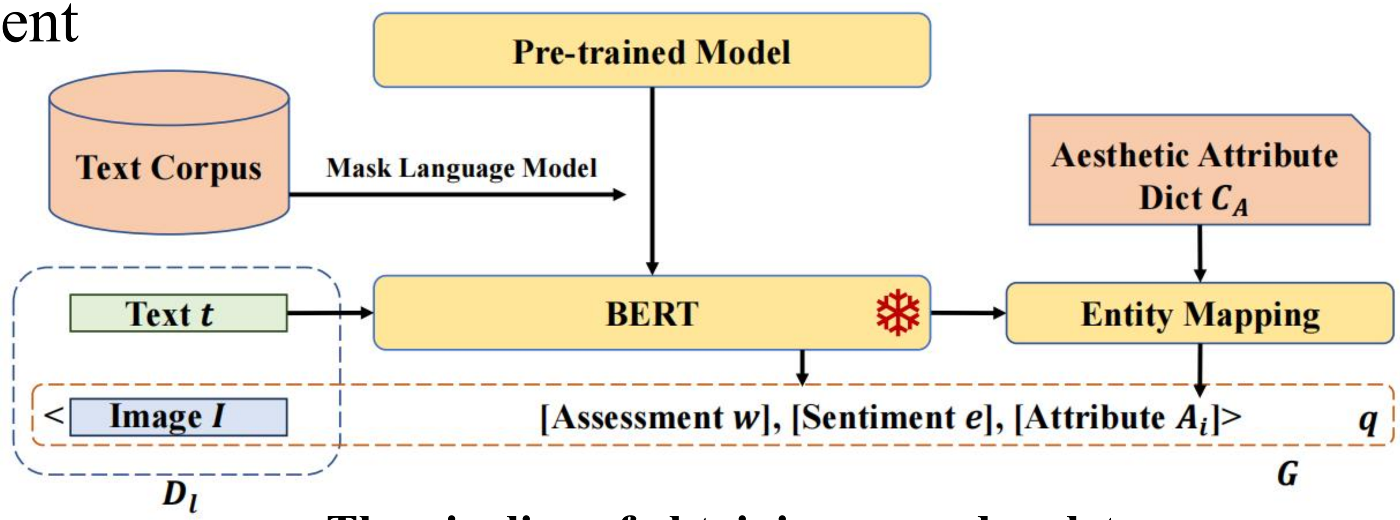
$$q = \langle \boldsymbol{I}, w, A_i, e \rangle$$

fig:

:Very simple, effective shot.

Obtained quadruplet:

[fig,effective,Use of Camera,Exposure & Speed,0.9998]



The pipeline of obtaining a quadruplet

- 2. Sample filtering and classification

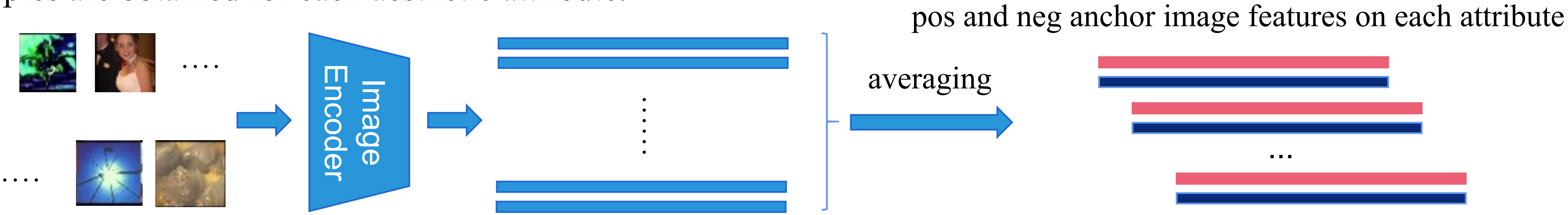
Set emotional score threshold and select images based on it.

Classify the selected images based on the positive and negative signs of polarity values and aesthetic attributes.

Methods - Internal Knowledge with Polarity

- 3. Obtain anchor images

Using the Image Encoder of the CLIP model to obtain the feature vectors of images for each category, and taking the average of them. From this, anchor plots representing positive and negative samples are obtained for each aesthetic attribute.



- 4. Calculate the score on a certain aesthetic factor according to formula (7)

$$s_i = \frac{e^{\boldsymbol{v}_s^{i,0}}}{e^{\boldsymbol{v}_s^{i,0}} + e^{\boldsymbol{v}_s^{i,1}}} \tag{7}$$

At Last, combined the wieghts from the external knowledge, we could obtain our final score by weighted summation of scores under different attributes

$$s = \frac{\sum_i^{N_a} s_i w_i}{N_a} \tag{8}$$

1. Background and Motivation

2. Methods

3. Experiments and Analysis

4. Conclusion and Future Work

Experiment - Dataset

➢AVA

- AVA dataset contains over 250,000 images.
- We use 19,929 images as the test split, following [10]

➢MMA (ours)

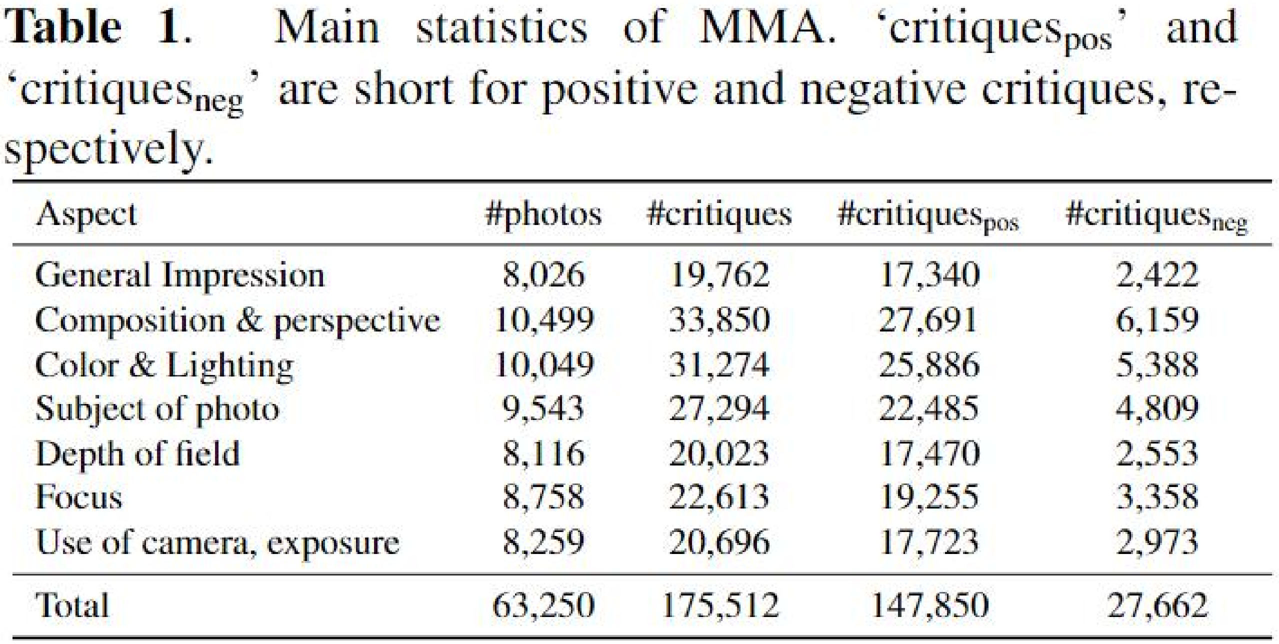


Table 1. Main statistics of MMA. ‘critiques$_{\text{pos}}$’ and ‘critiques$_{\text{neg}}$’ are short for positive and negative critiques, respectively.

Aspect	#photos	#critiques	#critiques$_{\text{pos}}$	#critiques$_{\text{neg}}$
General Impression	8,026	19,762	17,340	2,422
Composition & perspective	10,499	33,850	27,691	6,159
Color & Lighting	10,049	31,274	25,886	5,388
Subject of photo	9,543	27,294	22,485	4,809
Depth of field	8,116	20,023	17,470	2,553
Focus	8,758	22,613	19,255	3,358
Use of camera, exposure	8,259	20,696	17,723	2,973
Total	63,250	175,512	147,850	27,662

[10] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang, “Vila: Learning image aesthetics from user comments with vision-language pretraining,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10041– 10051.

Experiment - Metrics

- **We evaluate our KZIAA with some supervised competitors and zero-shot competitors. Our KZIAA outperforms many zero-shot competitor**

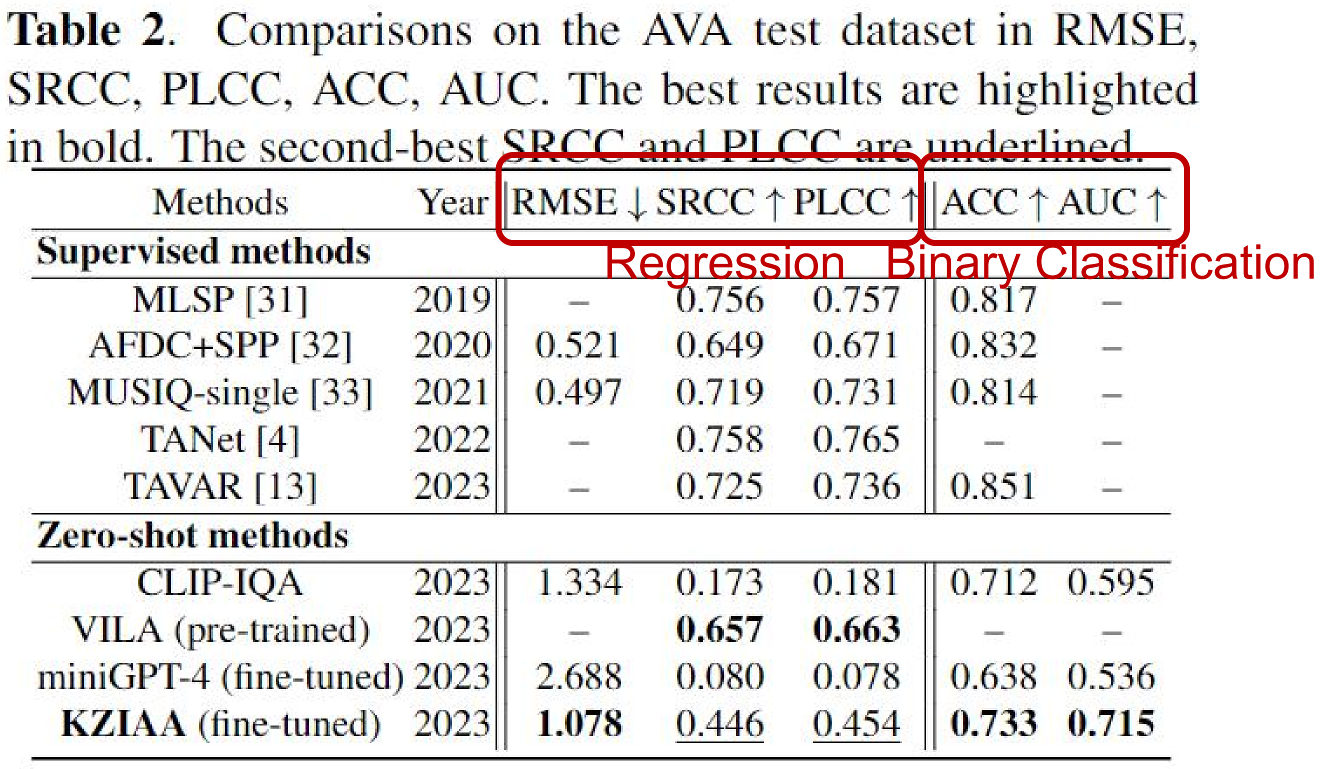


Table 2. Comparisons on the AVA test dataset in RMSE, SRCC, PLCC, ACC, AUC. The best results are highlighted in bold. The second-best SRCC and PLCC are underlined.

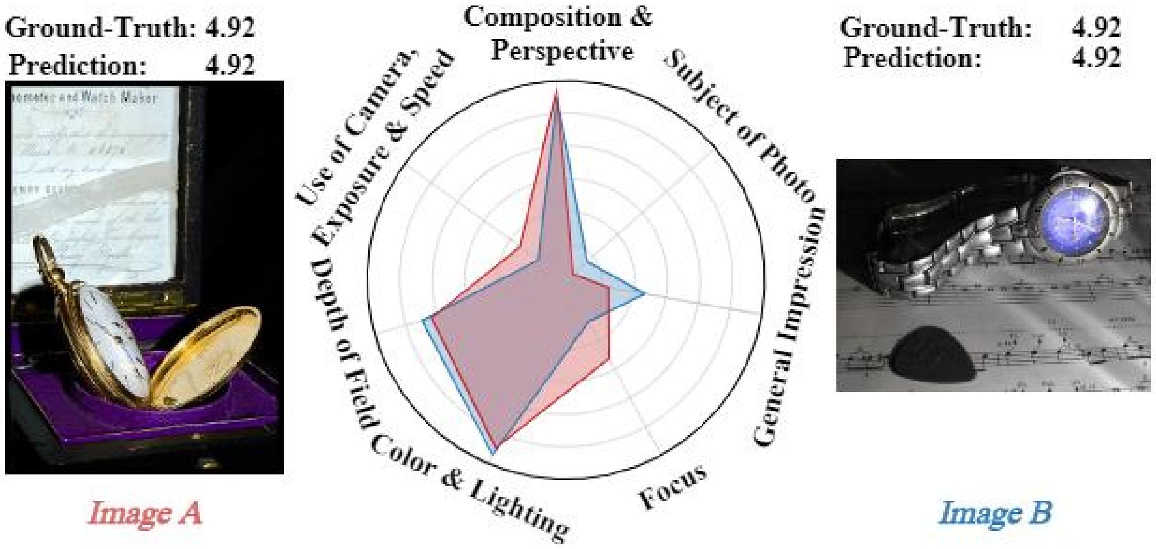
Regression

Binary Classification

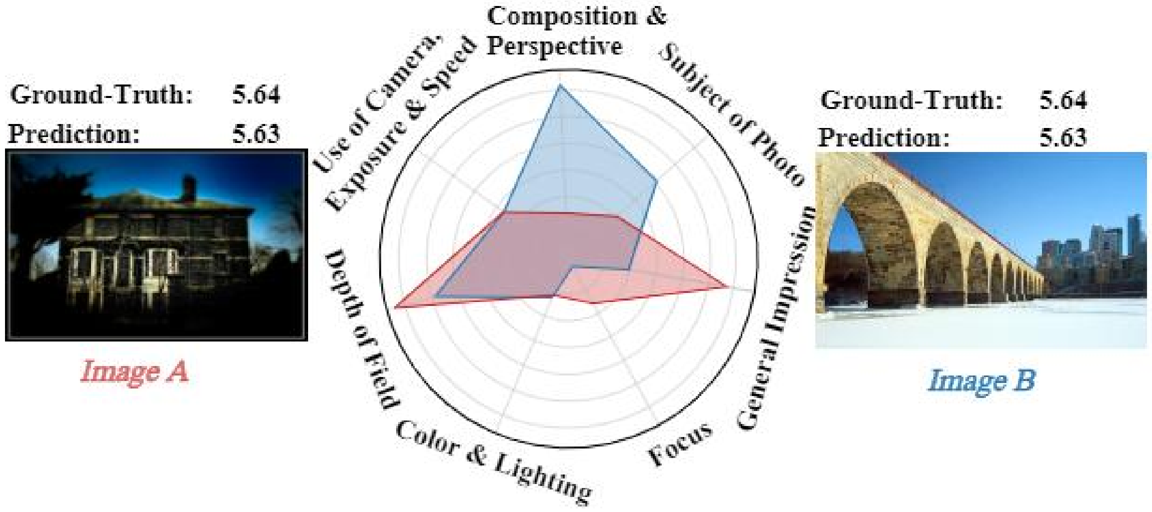
Methods	Year	RMSE ↓	SRCC ↑	PLCC ↑	ACC ↑	AUC ↑
Supervised methods						
MLSP [31]	2019	–	0.756	0.757	0.817	–
AFDC+SPP [32]	2020	0.521	0.649	0.671	0.832	–
MUSIQ-single [33]	2021	0.497	0.719	0.731	0.814	–
TANet [4]	2022	–	0.758	0.765	–	–
TAVAR [13]	2023	–	0.725	0.736	0.851	–
Zero-shot methods						
CLIP-IQA	2023	1.334	0.173	0.181	0.712	0.595
VILA (pre-trained)	2023	–	**0.657**	**0.663**	–	–
miniGPT-4 (fine-tuned)	2023	2.688	0.080	0.078	0.638	0.536
KZIAA (fine-tuned)	2023	**1.078**	0.446	0.454	**0.733**	**0.715**

Experiment - Analysis

- ***We can analyze aesthetic attributes first before obtaining the final score.***



(a)



(b)

Experiment - Ablation

- ***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.***

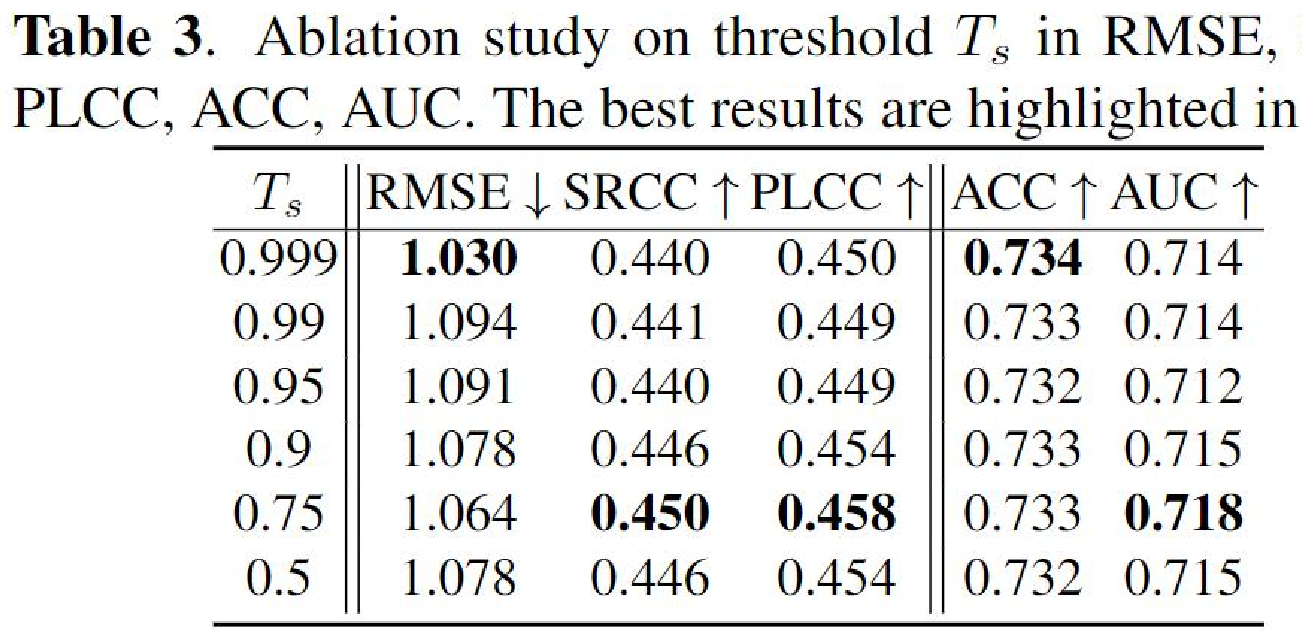


Table 3. Ablation study on threshold T_s in RMSE, SRCC, PLCC, ACC, AUC. The best results are highlighted in bold.

T_s	RMSE ↓	SRCC ↑	PLCC ↑	ACC ↑	AUC ↑
0.999	**1.030**	0.440	0.450	**0.734**	0.714
0.99	1.094	0.441	0.449	0.733	0.714
0.95	1.091	0.440	0.449	0.732	0.712
0.9	1.078	0.446	0.454	0.733	0.715
0.75	1.064	**0.450**	**0.458**	0.733	**0.718**
0.5	1.078	0.446	0.454	0.732	0.715

Conclusion

- 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.**

Future Work

- ***How to teach LMMs for Visual Scoring via customized sentence-level features?***
- ***The following prompts seem not enough***
  - Can you evaluate the aesthetics of the image?
  - Rate the aesthetics of this picture.
  - How is the aesthetics of this image?
  - Can you rate the aesthetics of this picture?
  - Please evaluate the aesthetics of the image.
  - Please tell the image quality in terms of aesthetics.

Future Work

- ***How to teach LMMs for Visual Scoring via customized sentence-level features?***
- ***We will try in the following directions:***
  - We will design sentence-level descriptions of image aesthetic attributes and connect them to different rating levels
  - We will design sentence-level descriptions of personality, rather than Big-Five code.
  - We will explore multi-modal feature representation of attributes, rather than uni-modal representation.

Thank you!

My Blog QR!