KEEP KNOWLEDGE IN PERCEPTION: ZERO-SHOT IMAGE AESTHETIC ASSESSMENT

Guolong Wang[1*†] *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. First, we use prompt tuning to get a unique prompt for each aesthetic attribute as external knowledge. Second, we construct a multi-modal knowledge graph using image aesthetic critiques and leverage image relations in the graph as internal correlations. Specifically, we obtain aesthetic attribute representations from pre-trained models via prompt learning, then select anchor images on specific aesthetic attributes by sentiment polarity, computing aesthetic scores. Notably, annotated scoring information is not used in the process. Experiments show that our zero-shot approach performs similarly to supervised methods using only a small knowledge graph.

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

1. INTRODUCTION

Image aesthetic assessment (IAA) is a significant task within the field of computer vision [1]. The feasibility of this task has been progressively augmented by advancements in deep learning methods based on large-scale databases [2]. 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 [3]. 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 [4] has achieved 91.93% [5] accuracy on the CUHKPQ dataset [6].

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 a crucial point for a practical method [7]. The pattern of correlations between image features and aesthetic scores is rather heterogeneous across different datasets [8]. Secondly, the annotated single score in current databases proved insufficient to meet the evolving needs. This drove the demand for zero-shot image aesthetic assessment.

Knowledge is crucial in a zero-shot setting. Semantic information is always used as knowledge, such as aesthetic attributes [9]. Some methods designed a multi-task framework by introducing a related task, such as emotion prediction [10]. Others designed network structures for a specific attribute [11]. Except for aesthetic attributes, some methods formulate external knowledge (e.g., object information [12] and image critiques [13]) and internal knowledge (e.g., self-supervision [14]).

In this work, we propose an **K**nowledge-enhanced **Z**ero-shot **I**mage **A**esthetic **A**ssessment (KZIAA) framework to address the limitations. Image aesthetic assessment is a subjective task where more semantic information than a score is useful [15, 16]. Consequently, attention shifted towards the finer-grained aspects of the aesthetic assessment of images [17, 18]. Thus, assessing different aspects of an image's aesthetics became increasingly valuable, desiring external knowledge. It can offer comprehensive information and enhance the generalization ability of models [19]. 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 omit the score annotation by employing a multimodal knowledge graph to enrich the aesthetic representation. Thus, image aesthetics combines external knowledge and internal relationships. The framework has three main novelties:

1) We expand the current AVA dataset to construct a multi-modal knowledge graph of image aesthetics.

2) We construct a continuous prompt with a pre-trained model and use transformer as a decoder to address the inference problem for new images.

3) We incorporate external knowledge and internal rela-

*Equal contribution. †Corresponding Author. This work was supported by "the Fundamental Research Funds for the Central Universities" in UIBE (22QN01).

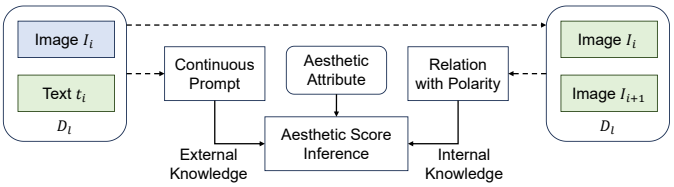

Fig. 1. Overview of our research design. It contains three components: (1) continuous prompt, (2) relation with polarity, and (3) aesthetic score inference. The white rounded rectangle denotes data and white rectangle denotes a component. The blue and green rectangle denotes image and text, respectively. The dashed arrow means data input/output to the model, and the arrow means data flow in the model.

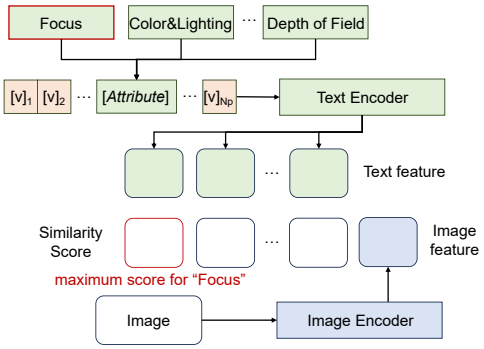

Fig. 2. Prompt tuning pipeline. Through prompt tuning, we obtain a unique context for each attribute.

tionship 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 $\boldsymbol{I}$, $\mathcal{M}$ can predict the aesthetic score s for this image:

$$s = \mathcal{M}(\boldsymbol{I}) \tag{1}$$

where $\boldsymbol{I} \in \mathbb{R}^{w \times h}$, 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 $\hat{s}$. Instead, we use a collected large-scale vision-language corpus $\mathcal{D}_l = \{(I_i, t_i)\}_{i=1}^{N_d}$ as external knowledge base and the relationship between images $\mathcal{D}_r = \{(I_i, I_{i+1})\}_{i=1}^{N_i-1}$ as internal knowledge. N_m and N_s is the number of multi-modal pairs and images, respectively. Following [17], we have seven aesthetic attributes for images. Let N_a to denote the number of attributes. $N_a = 7$ in this setting.

3. IMAGE AESTHETIC ASSESSMENT DESIGN

Our research design comprises three components: (1) External knowledge for aesthetic attributes with continuous prompt, (2) Internal knowledge for 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 prompt 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." [20]. However, the handcrafted templates are always sub-optimal, hindering the generalization capability. Thus, we utilize prompt tokens $[\boldsymbol{v}]_i, i \in [1, N_p]$ that will be optimized to formulate the continuous prompt templates $\boldsymbol{F}(Attribute)$ 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.

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

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 $\mathcal{D}_l$ 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 [21].

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 $< \cdot >$ commonly used in the literature. The probability of predicting that I is attribute A_i is:

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

where τ is the temperature parameter.

We use the cross entropy loss for attribute classification:

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

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

We get a unique prompt $\boldsymbol{F}(A_i)$ for each attribute through prompt tuning. Different from [20], 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.

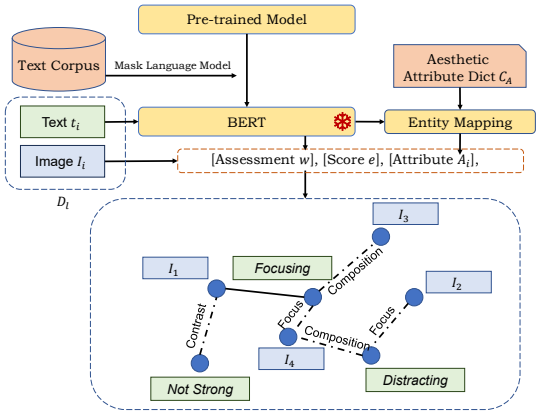

Fig. 3. The pipeline of building knowledge graph G.

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. [22]. Through comparison, they can shrink the estimation uncertainty due to the implicit knowledge of the relationship between two images.

Motivated by these observations, we also adopt the comparison with aesthetic representative images (aka anchor images) to build internal knowledge. As shown in Fig. 3, we further summarize a heterogeneous multi-modal knowledge graph G, where we have two types of nodes: (1) image $\boldsymbol{I}$ and (2) aesthetic assessment word w. For the word w, we extract it from a critique on attribute A_i for the corresponding image. The relation between node $\boldsymbol{I}$ and w is the attribute A_i. For a triplet $< \boldsymbol{I}, A_i, w >$, we use BERT [23] to measure the sentiment polarity score e of the critique.

We express the relationship between two images based on aesthetic assessment and sentiment polarity regarding different aesthetic attributes. It makes the anchor images contain varying aesthetic information, providing rich reference aesthetic knowledge in a contrasting perspective. Compared to the knowledge base in [24], our knowledge graph integrates multi-modal knowledge information and explicitly expresses the relationship between images.

3.3. Aesthetic Score Inference

In this secion, we make aesthetic score inference based on the external knowledge of attribute $\boldsymbol{F}(Attribute)$ and internal knowledge in knowledge graph G. The whole process is split into two phases: (1) knowledge embedding and (2) aesthetic decision making.

Firstly, we select anchor images with positive and negative polarity, denoted as $\mathcal{I}_p$ and $\mathcal{I}_n$. Specifically, we select images whose absolute value of sentiment polarity score is larger than threshold T_s from G. Then we calculate the similarity between the test image and anchor images in $\mathcal{I}_p$ and $\mathcal{I}_n$ on each attribute, respectively. Thus, given a test image I, we can get a similarity vector $\boldsymbol{v}_s \in \mathbb{R}^{2\times N_a}$. 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 Eqn. 5, 6, and 7.

$$s_i = \frac{e^{\boldsymbol{v}_s^{i,0}}}{e^{\boldsymbol{v}_s^{i,0}} + e^{\boldsymbol{v}_s^{i,1}}} \tag{5}$$

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

$$s = \frac{\sum_i^{N_a} s_i w_i}{N_a} \tag{7}$$

Table 1. Main statistics of MMA. 'critiques$_{pos}$' and 'critiques$_{neg}$' is short for positive and negative critiques, respectively.

Aspect	#photos	#critiques	#critiques$_{pos}$	#critiques$_{neg}$
General Impression	723	955	881	74
Composition & perspective	1,091	1,666	1,503	163
Color & Lighting	1,018	1,556	1,428	128
Subject of photo	878	1,248	1,123	125
Depth of field	738	983	916	67
Focus	799	1,090	1,002	88
Use of camera, exposure	749	1,008	932	76
Total	5,996	7,551	7,785	721

4. EXPERIMENT

4.1. Database

AVA [25]. It has over 250,000 images [1]. We use the mean value of all ratings for an image as its ground-truth aesthetic score. In our zero-shot setting, we only use the test dataset.

Multi-Modal Aesthetic (MMA) Database. We collect image aesthetic critiques for images in AVA to build the multi-modal knowledge graph: (1) We collect critiques on the original website following [13] and use critiques in [16] as a supplement. (2) We clean the collected data by eliminating three types of sentences: incomplete sentences (e.g., 'This is a'), ambiguous sentences (e.g., 'Good boy'), and sentences only containing a score. (3) We classify the critiques into seven aspects according to keywords summarized by professional photographers and calculate the emotional polarity of each critique by BERT [23]. We show MMA's statistics in Table 1. Note that our MMA only contains images in the training set of AVA and has no access to ground-truth scores.

4.2. Evaluation Metrics

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

RMSE, SRCC, LCC: RMSE measures the difference between predicted values and actual observed values. SRCC and

[1] https://www.dpchallenge.com/

Table 2. Overall comparisons on the AVA test dataset in LCC, SRCC, RMSE, ACC, AUC.

Methods	Year	RMSE ↓	SRCC ↑	LCC ↑	ACC ↑	AUC ↑
Supervised methods						
MLSP [29]	2019	–	0.756	0.757	0.817	–
AFDC+SPP [30]	2020	0.521	0.649	0.671	0.832	–
MUSIQ-single [31]	2021	0.497	0.719	0.731	0.814	–
TANet [3]	2022	–	0.758	0.765	–	–
TAVAR [11]	2023	–	0.725	0.736	0.851	–
proposal-based methods						
CLIP-IQA	2023	1.334	0.173	0.181	0.712	0.595
miniGPT-4	2023	2.471	0.043	0.038	0.711	0.560
KZIAA	2023	**1.078**	**0.446**	**0.454**	**0.733**	**0.715**

LCC refer to Spearman's rank and Pearson correlation coefficients between the predicted and ground-truth mean scores, respectively. These two metrics measure models' ability of ranking different images on aesthetic score.

ACC, AUC: We can reduce IAA to a binary classification problem. The images whose score is above five is 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.

4.3. Implementation Details

We use CLIP [26] as the basic vision-language pre-trained model. Specifically, we use ViT-B/32 for the image encoder and Transformer for the text encoder. For fine-tuning the continuous prompt, we insert the context tokens before aesthetic attributes. We set $N_p = 16$ and the size of $[Attribute]$ 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 [27] with a learning rate of $1e{-}3$. The BERT we used in knowledge graph construction is pre-trained following [23]. 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 Python 3.7 on 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: CLIP-IQA [20] and miniGPT-4 [28]. The results summarized in Table 2 show that our KZIAA outperforms zero-shot baselines. Notably, it has the potential to catch up with supervised methods, as we only use a small knowledge graph containing 5,996 images.

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

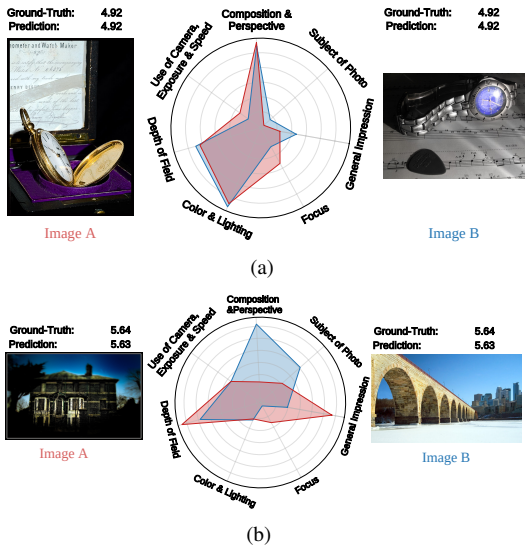

Fig. 4. (a) The two images' content, lighting, and composition were highly similar. Our model gives predictions identical to the ground-truth scores and has hugely consistent attention to attributes. (b) Our model gives the same score for the two images. However, it focuses on General Impression and Depth of Field in Image A, while focusing on the Composition and Depth of Field in Image B.

Table 3. Ablation study on threshold T_s in LCC, SRCC, RMSE, ACC, AUC.

T_s	RMSE ↓	SRCC ↑	LCC ↑	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

Qualitative Analysis of prompts $\boldsymbol{F}(Attribute)$. We show an exemplar of continuous prompts by visualizing w_i(Eqn. 6) in Fig. 4. These weights contribute to final score calculation.

5. 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 image relationships. Firstly, we obtain a unique context for each aesthetic attribute by prompt tuning. Subsequently, we build a knowledge graph and utilize sentiment polarity to select anchor image nodes in the graph. Finally, we estimate the score considering the information of different attributes. Experiment results indicate the superiority of the proposed method to zero-shot baselines and the potential to approach supervised methods.

References

[1] Luming Zhang, Yue Gao, Roger Zimmermann, Qi Tian, and Xuelong Li, "Fusion of multichannel local and global structural cues for photo aesthetics evaluation," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1419–1429, 2014.

[2] Yubin Deng, Chen Change Loy, and Xiaoou Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017.

[3] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming, "Rethinking image aesthetics assessment: Models, datasets and benchmarks," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 2022, pp. 942–948.

[4] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[5] Zhe Dong, Xu Shen, Houqiang Li, and Xinmei Tian, "Photo quality assessment with dcnn that understands image well," in *MultiMedia Modeling: 21st International Conference, MMM 2015, Sydney, NSW, Australia, January 5-7, 2015, Proceedings, Part II 21*. Springer, 2015, pp. 524–535.

[6] Xiaoou Tang, Wei Luo, and Xiaogang Wang, "Content-based photo quality assessment," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[7] Ziyue Wu, Xi Chen, and Zhaoxing Gao, "Bayesian non-parametric method for decision support: Forecasting online product sales," *Decision Support Systems*, p. 114019, 2023.

[8] Yangyang Shu, Qian Li, Lingqiao Liu, and Guandong Xu, "Semi-supervised adversarial learning for attribute-aware photo aesthetic assessment," *IEEE Transactions on Multimedia*, 2021.

[9] Yangyang Shu, Qian Li, Lingqiao Liu, and Guandong Xu, "Privileged multi-task learning for attribute-aware aesthetic assessment," *Pattern Recognition*, vol. 132, pp. 108921, 2022.

[10] Jun Yu, Chaoran Cui, LeiLei Geng, Yuling Ma, and Yilong Yin, "Towards unified aesthetics and emotion prediction in images," in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 2526–2530.

[11] Leida Li, Yipo Huang, Jinjian Wu, Yuzhe Yang, Yaqian Li, Yandong Guo, and Guangming Shi, "Theme-aware visual attribute reasoning for image aesthetics assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[12] Jingwen Hou, Henghui Ding, Weisi Lin, Weide Liu, and Yuming Fang, "Distilling knowledge from object classification to aesthetics assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7386–7402, 2022.

[13] Guolong Wang, Junchi Yan, and Zheng Qin, "Collaborative and attentive learning for personalized image aesthetic assessment.," in *IJCAI*, 2018, pp. 957–963.

[14] Jan Pfister, Konstantin Kobs, and Andreas Hotho, "Self-supervised multi-task pretraining improves image aesthetic assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 816–825.

[15] Yueying Kao, Ran He, and Kaiqi Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1482–1495, 2017.

[16] Zhipeng Zhong, Fei Zhou, and Guoping Qiu, "Aesthetically relevant image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 3733–3741.

[17] Kuang-Yu Chang, Kung-Hung Lu, and Chu-Song Chen, "Aesthetic critiques generation for photos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3514–3523.

[18] Daniel Vera Nieto, Luigi Celona, and Clara Fernandez Labrador, "Understanding aesthetics with language: A photo critique dataset for aesthetic assessment," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34148–34161, 2022.

[19] Michal Kucer, Alexander C Loui, and David W Messinger, "Leveraging expert feature knowledge for predicting image aesthetics," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5100–5112, 2018.

[20] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy, "Exploring clip for assessing the look and feel of images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 2555–2563.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *ECCV*, 2016, pp. 662–679.

[23] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[24] Leida Li, Tianwu Zhi, Guangming Shi, Yuzhe Yang, Liwu Xu, Yaqian Li, and Yandong Guo, "Anchor-based knowledge embedding for image aesthetics assessment," *Neurocomputing*, vol. 539, pp. 126197, 2023.

[25] Naila Murray, Luca Marchesotti, and Florent Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2408–2415.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.

[27] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[29] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9375–9383.

[30] Qiuyu Chen, Wei Zhang, Ning Zhou, Peng Lei, Yi Xu, Yu Zheng, and Jianping Fan, "Adaptive fractional dilated convolution network for image aesthetics assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14114–14123.

[31] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang, "Musiq: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5148–5157.