Deep Learning for Image Aesthetic Assessment: A Comprehensive Survey

Guangao Wang

***Abstract*—Image aesthetic evaluation aims to automatically predict the aesthetic quality of an image or people's aesthetic preferences through computers. With the development of deep learning, this field has made remarkable progress in the past decade. This paper systematically reviews the image aesthetic evaluation methods based on deep learning, including task definition, common datasets, model method classification, loss function and evaluation metrics, as well as current challenges and future development directions. We discuss the early methods of using convolutional neural networks (CNNs) for aesthetic rating classification/regression, as well as improved methods that integrate multi-task learning, attention mechanisms, image semantic understanding, and contrastive learning. At the same time, we introduce recent new methods based on the Transformer architecture (such as ViT, CLIP, etc.), and compare the differences between end-to-end training and feature fusion strategies. Finally, the challenges faced by this field (such as aesthetic subjectivity, cross-domain generalization, user personalization, multimodal fusion, etc.) and future development trends are discussed.**

***Index Terms*—Image Aesthetic Assessment, Deep Learning, Convolutional Neural Networks, Vision Transformers.**

I. Introduction

In human visual experience, the aesthetic evaluation of images is of great significance and is widely used in the fields of photography selection, image retrieval, social media recommendation, etc. The image aesthetic evaluation task aims to develop computational models to simulate human evaluation of the aesthetic quality of images. Since aesthetic evaluation is obviously subjective, different people may have different aesthetic preferences for the same image, which makes this task extremely challenging [1]. Early studies mainly used manually designed image features and traditional machine learning methods to distinguish high-quality and low-quality photos [2]. However, these methods have limitations in feature expression and generalization capabilities.

In recent years, deep learning methods (especially convolutional neural networks) have made breakthroughs in the field of computer vision, which has also promoted the research of image aesthetic evaluation into a new stage [3]. Deep models can automatically learn multi-level representations of images, including low-level visual features (such as color, composition) and high-level semantic information, so as to better predict human aesthetic evaluation. Many researchers have proposed various models based on deep learning and achieved excellent results on large-scale aesthetic datasets. At the same time, some review works have summarized this field. Based on the existing work, this paper will systematically sort out the deep learning methods for image aesthetic evaluation.

We first formally define the image aesthetic evaluation problem and its task objectives, and then introduce commonly used datasets and their characteristics. Next, we review deep learning models by method category, including classic CNN-based methods, improved methods that integrate multi-task and attention mechanisms, new Transformer-based methods, and comparisons between end-to-end and feature fusion strategies. Finally, we discuss current challenges (such as subjectivity, cross-domain generalization, and personalization) and look forward to future development directions.

Manuscript received May 18, 2025

Guangao Wang, School of Software, Tiangong University, Tianjin, China

II. Dataset Overview

Publicly available large-scale datasets have promoted the training and evaluation of image aesthetic assessment models. Below we introduce several commonly used image aesthetic assessment datasets and their characteristics.

A. AVA

The AVA [4] dataset was created by Naila Murray et al. and is one of the most widely used large-scale aesthetic evaluation datasets. AVA contains about 250,000 images, which are mainly from the photography enthusiast website DPChallenge. Each image is rated by a large number of users (1 to 10 points). The average score of each image is usually used as the true value label of its aesthetic evaluation. AVA's ratings are widely distributed, and many images have intermediate scores (between high and low), representing "ambiguous" cases with medium aesthetic quality. AVA also provides 14 style attribute labels (such as "black and white", "macro", etc.) and more than 60 semantic category labels for some images, which can be used to study the relationship between aesthetics and attributes. AVA is commonly used for two evaluation tasks:

(1) Binary classification: images with an average score above a certain threshold (such as 5 or 5.5) are considered to have high aesthetic quality, and those below the threshold are considered to have low quality, and then the classification accuracy is calculated;

(2) Regression/correlation: directly predict the average score, and use the ranking correlation coefficient to evaluate the model's fit to the score ranking.

B. CUHK-PQ

CUHK-PQ [5] is provided by the Chinese University of

Hong Kong, China, and contains 17,690 images. These images are collected from DPChallenge and amateur photographers and are smaller in scale than AVA. Each image in CUHK-PQ is manually labeled as "high quality" or "low quality" and divided into 7 types according to the scene (such as animals, plants, still life, architecture, landscape, portrait, night scene). This dataset emphasizes binary classification tasks and often uses a random division of 50% training and 50% testing, or 5-fold cross validation. Due to the clear scene labels, some studies use scene-by-scene training or fuse scene information in the model to improve aesthetic classification performance.

C. AADB

AADB [6] was built by Kong et al. and released in 2016. It collects about 10,000 diverse photos from Flickr, with a balance of professional and amateur photos. Each picture is evaluated by multiple people with an aesthetic score (the average value is taken as the label) and 11 aesthetic attribute scores (such as composition, lighting, color, depth of field, etc.). AADB also records the anonymous ID of the rater, which can be used for personalized analysis. The official standard division is: 8,500 training, 500 verification, and 1,000 testing. Unlike AVA, the scoring scale of AADB is usually 1 to 5 points. The characteristics of AADB are that it provides auxiliary information of aesthetic attributes, supports multi-task learning (simultaneously predicting aesthetic scores and attributes) and ranking learning (the paper uses a network with ranking loss to learn relative aesthetic ranking).

D. Other datasets

As the research deepens, some special fields or larger-scale datasets have also emerged. For example, Yi et al. released an art photo aesthetics dataset BAID [7], focusing on the aesthetic evaluation of artistic style images to promote the performance of models in the art field. The general trend is that the scale of the dataset is getting larger and more diverse to cover images of different styles and fields, thereby improving the generalization ability of the model.

III. Deep Learning Models

Image aesthetic assessment methods based on deep learning can be roughly divided into the following categories: (1) Image aesthetic scoring classification or regression methods based on the convolutional neural network (CNN) architecture; (2) Models improved by multi-task learning, attention mechanism, semantic understanding, contrastive learning and other technologies; (3) Models based on the Transformer structure (such as directly using the ViT model or using the CLIP pre-trained model); (4) End-to-end training methods and methods based on feature fusion. The following introduces the development context and representative works of each type of method.

A. Rating classification and regression based on CNN

The introduction of convolutional neural networks (CNNs) was an important turning point in the field of image aesthetic assessment. Early researchers began to use CNNs to automatically learn features from images, replacing traditional hand-crafted features, which significantly improved the performance of aesthetic prediction. The work of Lu et al. [8] is one of the pioneers in this direction: they proposed a model to apply deep learning to image aesthetic scoring. This model uses a deep CNN to learn features directly from the original image and use these features as input to predict the aesthetic score or category of the image. Experiments show that compared with the traditional hand-crafted feature + classifier method, deep CNN can more effectively distinguish between high- and low-aesthetic images.

Subsequently, Lu et al. further proposed the Deep Multi-patch Aggregation Network [9]. The model randomly samples multiple local patches of the image, extracts local features separately through CNN, and then aggregates these features to predict the overall aesthetic score. In this way, multi-scale local information can be used to improve prediction accuracy without losing global information. Experiments at the time showed that the multi-patch strategy can improve the ability to capture aesthetic elements in complex scenes and make the model more sensitive to composition and local details.

Mai et al. proposed a composition-preserving deep aesthetic assessment model [10]. They realized that conventional CNN inputs require cropping or scaling images to a fixed size, which may destroy the original aspect ratio and composition of the image. To this end, they designed a special image cropping/pooling strategy to preserve the composition ratio of the original image as much as possible, and fed the processed image into CNN, reducing the loss of aesthetic information caused by deformation. This method achieved better performance on the AVA benchmark that year, proving the importance of composition information in aesthetic assessment.

With the evolution of CNN architecture, researchers have gradually tried deeper networks and pre-trained models. For example, Talebi and Milanfar proposed the famous NIMA model [3]. NIMA is based on powerful convolutional networks such as Inception-V2, and fine-tunes the ImageNet pre-trained weights to predict aesthetic evaluations. Unlike the previous output of only an average score or category, NIMA outputs an aesthetic score distribution (i.e., a probability distribution of 1-10 points), which can calculate the average aesthetic score of the image and the variance reflecting the uncertainty of the evaluation. They used Earth Mover's Distance (EMD) as the loss metric to measure the difference between the predicted distribution and the true score distribution to train the network. NIMA achieved leading performance on the AVA dataset at the time and was able to output meaningful rating uncertainty information.

In addition to the above representatives, there are many other explorations of CNN-based aesthetic evaluation models. For example: the A-Lamp network proposed by Ma et al. [11] uses adaptive layout-aware multi-branch CNN to better process photos with different layouts; Kao et al. combines CNN features and semantic information to improve prediction; Jin et al. [12] extracts multi-level features for aesthetic prediction; and some models that integrate residual networks (ResNet), VGG and other architectures. This series of studies has established the dominant position of CNN in aesthetic evaluation and proved that end-to-end deep feature learning is effective in capturing complex aesthetic factors.

B. Improvement methods

Kong et al. [6] jointly predicted the aesthetic score and 11 attribute scores in AADB, and used attribute prediction as an auxiliary task to regularize the main task. Zhu et al. [13] adopted a meta-learning framework to achieve personalized aesthetic evaluation. Sheng et al. [14] proposed an attention-based multi-patch aggregation model to emphasize key parts by learning regional attention weights. Hou et al. [15] introduced attention at the object level, first detecting salient objects in the image and then aggregating object features. She et al. [16] constructed a hierarchical layout-aware graph convolutional network (Layout-aware GCN), dividing the image into regional nodes to build a map to simulate the influence of composition rules on aesthetics. Ko et al.'s [17] high-level attribute idea was also inherited in the deep era to capture semantic attributes such as "happy" and "natural". Sheng et al. [18] performed self-supervised pre-training on unlabeled data and then fine-tuned it on the aesthetic evaluation task to improve feature discrimination. The method based on the contrastive learning framework uses high/low aesthetic image pairs as positive and negative samples to train the model to learn the relative relationship in the embedding space [19]. The cross-modal features learned by the CLIP model through pre-training on image-text comparison have also been shown to have good transferability for aesthetic evaluation [20].

In summary, various improved methods have strengthened the model's understanding of aesthetic factors through multi-task, attention, semantic and contrastive learning techniques based on CNN, and are often used in combination to achieve the best performance.

C. Aesthetic evaluation method based on Transformer

In recent years, the Transformer architecture has achieved great success in the visual field (ViT model) and cross-modal field (CLIP model), and has also begun to be applied to image aesthetic evaluation tasks. Transformer's self-attention mechanism and powerful modeling capacity provide a new way to extract and fuse aesthetic features.

ViT [21] divides the image into patches of fixed size, regards each patch as a sequence token, and models the information of the entire image through the self-attention mechanism. Unlike CNN, which requires a fixed input size, ViT can theoretically process images of any size. However, in actual training, due to the limitations of position encoding and computational complexity, ViT usually uses cropped or scaled fixed-size images. This preprocessing will bring similar problems to the aforementioned CNN: destroying the original composition and details of the image, which may affect aesthetic judgment. Behrad et al. [22] proposed a method called Charm, which specifically solves the problem of ViT processing high-resolution full-size images in aesthetic evaluation. Charm uses a new tokenization strategy to preserve the aspect ratio and important high-resolution areas of the image, and only downsamples the minor parts, thereby encoding the image into a shorter sequence without cropping the key information. This method allows the pre-trained ViT to be directly used for aesthetic evaluation and achieve better performance: it improves the performance by about 8% compared to the standard ViT on multiple aesthetic and quality datasets. This proves that the Transformer combined with an appropriate high-resolution processing strategy can be competent for aesthetic evaluation and is superior to simple scaling schemes in composition sensitivity.

CLIP [23] is a cross-modal model proposed by OpenAI. After massive image-text pair comparison learning and training, it has powerful image representation capabilities. Researchers found that the image features extracted by CLIP contain rich style and semantic information, which is very suitable for the needs of aesthetic evaluation. Hentschel et al. [20] demonstrated the potential of CLIP in aesthetic evaluation through a series of experiments. First, they designed text prompts to allow CLIP to directly evaluate the beauty of images in a "zero-shot" way, such as calculating the matching degree between the text description "A beautiful photo" and the image, thereby obtaining an aesthetic score. Then, they fixed the image encoder of CLIP and only trained a layer of linear regression to map CLIP features to the AVA average score, which exceeded the performance of ImageNet pre-trained CNN features using the same method. Finally, they fine-tuned CLIP's image encoder for prediction on AVA, and only a few training iterations were required to achieve higher performance than fine-tuning ImageNet-CNN. These experiments show that compared with traditional classification pre-trained CNNs, CLIP's features are more suitable for aesthetic evaluation because CLIP is forced to learn high-level image attributes related to natural language descriptions during training, including subjective feelings, composition, and style.

In addition to using ViT and CLIP directly, there are also works that incorporate Transformer ideas into customized models. For example, He et al. proposed the EAT [24] (Enhancer for Aesthetics-Oriented Transformers) method, which combines visual Transformers with convolutional features to better predict aesthetics. Some studies use ViT in a meta-learning framework to achieve personalized aesthetic evaluation, adapting to the aesthetic preferences of different users by making small adjustments to the Transformer. Others explore the use of self-supervised Transformers for aesthetics, such as using the Transformer architecture for pre-training and then fine-tuning. Overall, Transformers provide a new modeling paradigm for image aesthetic evaluation: the self-attention mechanism helps capture the aesthetic associations of different regions within an image, and the general visual language knowledge brought by large-scale pre-training (such as CLIP) can be transferred to aesthetic tasks. As the visual Transformer model continues to mature, we have reason to believe that Transformer-based aesthetic evaluation methods will continue to make breakthroughs.

IV. Challenges and Future Directions

Although image aesthetic assessment has made great progress driven by deep learning, there are still many challenges to be solved. Combining the existing research trends, we believe that the following aspects are particularly important and may become the main development directions in the future:

A. Subjectivity and consistency in aesthetic evaluation

Beauty is a highly subjective concept, and different

individuals and cultural backgrounds may have very different evaluations of the same image. This leads to noise and inconsistency in data annotation: even after averaging, the "true value" aesthetic score of an image may have a large variance. This natural label noise makes model training difficult and may limit the upper limit performance of the model. Future research needs to better model the human uncertainty of aesthetic evaluation. On the one hand, consistency can be improved through more sophisticated annotation or filtering at the data level, such as recording the score distribution of each image, the background of the rater, and other information, and using statistical methods to evaluate the annotation reliability. On the other hand, uncertainty estimation can be introduced at the model level to output a predicted distribution or credible interval instead of a single value to characterize the uncertainty of the model's own aesthetic judgment. This helps the model reduce overconfidence when judging subjective difficult cases. There are also studies that set the upper limit of the model's performance by analyzing the correlation between human raters in order to clarify the room for improvement. In short, we should face up to the noise caused by aesthetic subjectivity and develop robust learning methods (such as noise-resistant loss, self-training, etc.) to improve the model's tolerance to label inconsistencies.

B. Generalization across domains and styles

Most current models are trained and tested on specific datasets, but aesthetic evaluation may vary in different image domains. For example, the aesthetic judgment standards for landscape photography, portrait photography, advertising images, and art paintings are not exactly the same. A model trained on AVA may not be able to evaluate the beauty of art paintings or illustrations well. This reflects the challenge of cross-domain generalization. In the future, more extensive and diverse training data (for example, covering various fields such as professional photography, mobile phone photos, and online pictures) or transfer learning/domain adaptation techniques are needed to enable the model to adapt to the aesthetic standards of new fields. Some studies have begun to try to adjust the model's scoring strategy for different categories of images through content labels, or to generate data from different domains through style transfer methods to enrich training. Recently emerged large-scale pre-trained models (such as CLIP) show promise in cross-domain generalization because they learn knowledge across diverse data. In the future, we can explore using these models as a basis and then do a small amount of tuning for specific aesthetic tasks to obtain a more universal aesthetic evaluation model. In addition, it is also necessary to consider the aesthetic differences in different cultural circles. Cross-cultural data collection and model calibration may be required to make the model output reasonable for users in different regions.

C. Personalized aesthetic assessment

An ideal aesthetic evaluation system should be able to take into account the personalized preferences of users. Most current models output scores based on the public aesthetic orientation, but everyone has different aesthetic tastes. For example, some people prefer strong and saturated colors, while others prefer minimalist compositions. Allowing the model to adjust its aesthetic evaluation based on the historical preferences of specific users can improve satisfaction with user-related tasks (such as album management and personalized recommendations). To this end, researchers have explored a variety of approaches: one is a personalized model, that is, training or fine-tuning an exclusive aesthetic evaluation model for each user, such as Ren et al. [25] by introducing user embedding vectors into the network to output different scores for different users; the other is meta-learning and few-shot learning, such as Zhu et al. [13] using a meta-learning framework to enable the model to quickly adjust to the user's aesthetic habits through a small amount of user feedback. There are also studies that use user interaction data (such as browsing and like records) to implicitly infer user preferences and integrate them into model decisions. Future aesthetic evaluation systems may need to learn user preferences online and continuously update models to truly achieve aesthetic judgments that vary from person to person. Of course, this also brings new challenges, including privacy and security issues in obtaining user preferences, and how to strike a balance between personalization and public evaluation.

D. Multimodal aesthetic assessment

The beauty of an image depends not only on the image itself, but is also often associated with its title, description, background music (for videos), and even social comments. For example, a photo accompanied by a beautiful poem description may enhance the viewer's subjective feelings; in social media, people's comments and likes on a picture also reflect its popularity. Therefore, combining images with other modal information for aesthetic evaluation is a direction worth exploring. Recent work has attempted to use text descriptions of images or user comments as auxiliary signals. Ke et al. [26] use visual-language pre-training to enable the model to extract aesthetic-related information from user comments, such as "perfect composition" or "messy colors" mentioned in the comments, thereby improving prediction accuracy. At the same time, cross-modal models such as CLIP have shown that language supervision can help models understand more abstract concepts. Future research can further use multimodal Transformers to jointly encode the visual features of an image with its related text (title, tags, comments) to output a more comprehensive aesthetic evaluation. This may be particularly useful for some scenes that are highly subjective and require semantic interpretation, such as art appreciation. The model can combine the text of the artwork description to judge its artistic value and beauty. Multimodal fusion can also be extended to aesthetic analysis in the audio and video fields, such as judging the aesthetic atmosphere of a film clip by combining the soundtrack. It should be noted that multimodal methods need to deal with information misalignment and noise between different modalities, but it undoubtedly provides new opportunities to improve the cognitive level of the model.

E. Other challenges and trends

In addition to the above-mentioned key points, there are some other aspects of image aesthetics evaluation that deserve attention. For example, the interpretability of the model: deep models are often black boxes, and we need a

more intuitive way to explain why the model gives a high score to a certain photo (such as by visualizing the attention area or generating human-friendly evaluation reasons), which is important for professional applications. Another example is real-time and efficiency: Real-time aesthetic scoring on mobile devices is valuable for photography assistance, which requires a lighter model or efficient reasoning method. Combination with image enhancement: Some studies have begun to integrate aesthetic evaluation into automatic image optimization (such as automatic cropping, filter recommendation), forming a closed loop so that the model can not only evaluate aesthetics but also improve aesthetics. Finally, with the development of generative AI, we may witness the emergence of generative models that can create highly aesthetic images, in which the evaluation model can be used as a discriminator or optimization target, which will further promote the improvement of the evaluation model.

V. CONCLUSION

Image aesthetic assessment, as a topic that connects subjective human aesthetics with objective image analysis, has made remarkable progress in the past decade. From the initial artificial feature classifier to today's complex models that can approximate human evaluation with the help of deep learning, we are gradually moving towards the goal of making machines "understand beauty". The introduction of deep learning models (CNN and Transformer) has greatly improved the performance of aesthetic assessment, among which convolutional neural networks are good at learning visual patterns, and self-attention models further capture global relationships. Techniques such as multi-task, attention, and contrastive learning make the model's understanding of beauty more in-depth and diverse.

However, we are also clearly aware that machine aesthetic assessment has not yet solved the core problem of subjectivity. Models sometimes have difficulty handling ambiguous cases between beauty and ugliness, and their adaptability to cross-domain data is also limited. Future research needs to pay more attention to the inclusiveness of data and algorithms to the diversity of human aesthetics. By integrating multimodal information, introducing personalization mechanisms, and leveraging large-scale cross-domain knowledge, we are expected to train more general and flexible aesthetic assessment models. In addition, combining aesthetic assessment with tasks such as generation and editing will open a new chapter in intelligent image creation. In general, image aesthetic assessment is a fascinating field that combines art and AI. With the development of deep learning and big data, we have reason to expect that it will make more exciting breakthroughs in the near future, bringing a qualitative leap in machine understanding of human aesthetics.

REFERENCES

[1] S. He, Y. Zhang, R. Xie, D. Jiang, and A. Ming, "Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, Vienna, Austria: International Joint Conferences on Artificial Intelligence Organization, Jul. 2022, pp. 942–948. doi: 10.24963/ijcai.2022/132.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying Aesthetics in Photographic Images Using a Computational Approach," A. Leonardis, H. Bischof, and A. Pinz, Eds., in Lecture Notes in Computer Science, vol. 3953. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 288–301. doi: 10.1007/11744078_23.

[3] H. Talebi and P. Milanfar, "NIMA: Neural Image Assessment," *IEEE Trans. on Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018, doi: 10.1109/TIP.2018.2831899.

[4] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2408–2415, Jun. 2012, doi: 10.1109/CVPR.2012.6247954.

[5] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *2011 International Conference on Computer Vision*, Nov. 2011, pp. 2206–2213. doi: 10.1109/ICCV.2011.6126498.

[6] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo Aesthetics Ranking Network with Attributes and Content Adaptation," B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., in Lecture Notes in Computer Science, vol. 9905. Cham: Springer International Publishing, 2016, pp. 662–679. doi: 10.1007/978-3-319-46448-0_40.

[7] R. Yi, H. Tian, Z. Gu, Y.-K. Lai, and P. L. Rosin, "Towards Artistic Image Aesthetics Assessment: a Large-scale Dataset and a New Method," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22388–22397, Jun. 2023, doi: 10.1109/CVPR52729.2023.02144.

[8] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "RAPID: Rating Pictorial Aesthetics using Deep Learning," in *Proceedings of the 22nd ACM international conference on Multimedia*, in MM '14. New York, NY, USA: Association for Computing Machinery, Nov. 2014, pp. 457–466. doi: 10.1145/2647868.2654927.

[9] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep Multi-patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 990–998. doi: 10.1109/ICCV.2015.119.

[10] L. Mai, H. Jin, and F. Liu, "Composition-Preserving Deep Photo Aesthetics Assessment," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 497–506. doi: 10.1109/CVPR.2016.60.

[11] S. Ma, J. Liu, and C. W. Chen, "A-Lamp: Adaptive Layout-Aware Multi-patch Deep Convolutional Neural Network for Photo Aesthetic Assessment," presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Jul. 2017, pp. 722–731. doi: 10.1109/CVPR.2017.84.

[12] X. Jin *et al.*, "Aesthetic Attributes Assessment of Images," in *Proceedings of the 27th ACM International Conference on Multimedia*, in MM '19. New York, NY, USA: Association for Computing Machinery, Oct. 2019, pp. 311–319. doi: 10.1145/3343031.3350970.

[13] H. Zhu, L. Li, J. Wu, S. Zhao, G. Ding, and G. Shi, "Personalized Image Aesthetics Assessment via Meta-Learning With Bilevel Gradient Optimization," *IEEE Transactions on Cybernetics*, vol. 52, no. 3, pp. 1798–1811, Mar. 2022, doi: 10.1109/TCYB.2020.2984670.

[14] K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang, and B.-G. Hu, "Attention-based Multi-Patch Aggregation for Image Aesthetic Assessment," in *Proceedings of the 26th ACM international conference on Multimedia*, in MM '18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 879–886. doi: 10.1145/3240508.3240554.

[15] J. Hou, S. Yang, W. Lin, B. Zhao, and Y. Fang, "Learning Image Aesthetic Assessment from Object-level Visual Components," Apr. 04, 2021, *arXiv*: arXiv:2104.01548. doi: 10.48550/arXiv.2104.01548.

[16] D. She, Y.-K. Lai, G. Yi, and K. Xu, "Hierarchical Layout-Aware Graph Convolutional Network for Unified Aesthetics Assessment," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 8471–8480. doi: 10.1109/CVPR46437.2021.00837.

[17] K. Ko, J.-T. Lee, and C.-S. Kim, "PAC-Net: Pairwise Aesthetic Comparison Network for Image Aesthetic Assessment," *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 2491–2495, Oct. 2018, doi: 10.1109/ICIP.2018.8451621.

[18] K. Sheng *et al.*, "Revisiting Image Aesthetic Assessment via Self-Supervised Feature Learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Apr. 2020, pp. 5709–5716. doi: 10.1609/aaai.v34i04.6026.

[19] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep Multi-patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 990–998. doi: 10.1109/ICCV.2015.119.

[20] S. Hentschel, K. Kobs, and A. Hotho, "CLIP knows image aesthetics," *Front. Artif. Intell.*, vol. 5, Nov. 2022, doi: 10.3389/frai.2022.976235.

[21] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing*

Systems, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 6000–6010.

[22] F. Behrad, T. Tuytelaars, and J. Wagemans, "Charm: The Missing Piece in ViT fine-tuning for Image Aesthetic Assessment," May 15, 2025, *arXiv*: arXiv:2504.02522. doi: 10.48550/arXiv.2504.02522.

[23] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 26, 2021, *arXiv*: arXiv:2103.00020. doi: 10.48550/arXiv.2103.00020.

[24] S. He, A. Ming, S. Zheng, H. Zhong, and H. Ma, "EAT: An Enhancer for Aesthetics-Oriented Transformers," in *Proceedings of the 31st ACM International Conference on Multimedia*, in MM '23. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 1023–1032. doi: 10.1145/3581783.3611881.

[25] J. Ren, X. Shen, Z. Lin, R. Mech, and D. J. Foran, "Personalized Image Aesthetics," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 638–647. Accessed: May 18, 2025. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Ren_Personalized_Image_Aesthetics_ICCV_2017_paper.html

[26] J. Ke, K. Ye, J. Yu, Y. Wu, P. Milanfar, and F. Yang, "VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10041–10051, Jun. 2023, doi: 10.1109/CVPR52729.2023.00968.