https://doi.org/10.2352/lim.2024.5.1.1

Advances in Image Aesthetics Assessment: Concepts, Methods, and Applications

Luigi Celona and Simone Bianco
Department of Informatics, Systems and Communication, University of Milano-Bicocca, viale Sarca, 336 Milano, Italy
{first_name.last_name@unimib.it}

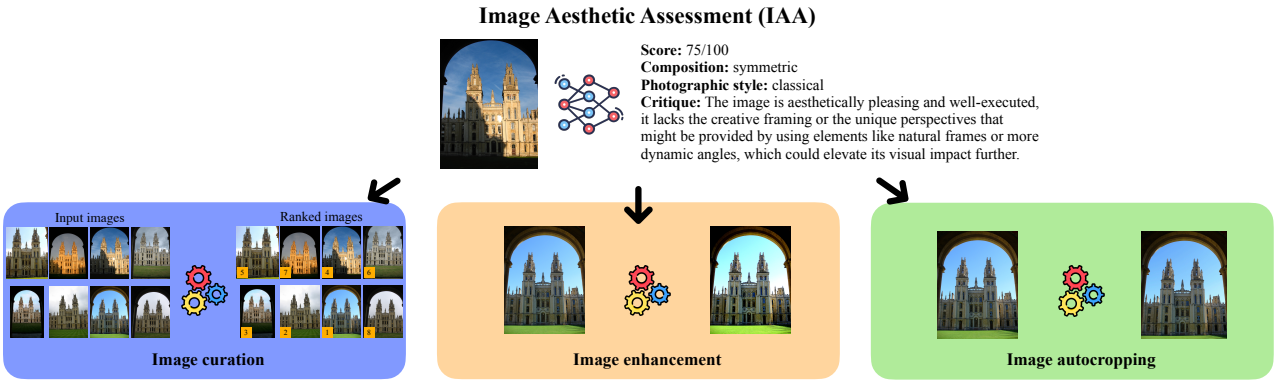

Figure 1: Image Aesthetic Assessment (IAA) automates the evaluation and prediction of an image's visual appeal using algorithms that generate aesthetic scores, criteria tags, or descriptive critiques. This technology is crucial for applications in content curation and photo editing, where rapid and automated judgment of visual quality is needed.

Abstract

Image Aesthetic Assessment (IAA) has attracted increasing attention recently but it is still challenging due to its high abstraction and complexity. In this paper, recent advancements in IAA are explored, emphasizing the goal, complexity, and critical role this task plays in improving visual content. Insights from our recent studies are combined to present a unified perspective on the state of IAA, focusing on methods relying on the use of genetic algorithms, language-based understanding, and composition-attribute guidance. These methods are examined for their potential in practical applications like content selection and quality enhancement, such as autocropping. The discussion concludes with an overview of the challenges and future directions in this field.

Introduction

Recent technological advancements in image and video acquisition, coding, and communication have democratized access to high-quality multimedia content, leading to an exponential increase in the volume of visual data shared on social platforms. Notably, 2.1 billion photos are uploaded daily on Facebook, and YouTube sees 2.4 million videos uploaded every day as of January 2024. This proliferation of visual media has significant social implications, especially in the realm of marketing where the valuation of content often hinges on user interactions such as "likes" [27]. These developments have underscored the necessity of automating the prediction of viewer preferences, driving advancements in computational aesthetics to understand and anticipate the types of images or videos that appeal to human observers through algorithmic means.

The mechanisms that drive image preference are diverse and complex, encompassing elements of interestingness [15], beauty, and memorability [1, 18]. These aspects are often intertwined and challenging to separate, posing significant challenges for the collection of unbiased datasets and the study of image aesthetics. Predicting visual preferences is not only economically valuable for applications such as advertising, personal photo management, and content retrieval but also crucial for enhancing our understanding of human perception.

Computational aesthetics, as defined by Neumann *et al.*, seeks to emulate human-like aesthetic judgments through algorithms that provide measurable and applicable outcomes [25]. The goal of Image Aesthetic Assessment (IAA) is to numerically quantify the visual appeal of an image, often supplemented by information on style, composition, or even a textual critique justifying the score (see Figure 1). IAA algorithms find application in various areas, including image cropping [5, 14], color and composition enhancement [11, 34], and personalized recommendations [8, 29, 17].

In this paper, we synthesize our findings from three distinct approaches to provide a comprehensive understanding of the current state of IAA:

- Genetic algorithm-based feature combination: Exploits genetic algorithms to assess the aesthetics of images containing faces through the combination of deep features [7].
- Composition and style attributes guidance: Explores how composition and style attributes can guide aesthetic quality predictions [6].
- Language-based aesthetics understanding: Integrates Natural Language Processing (NLP) with computer vision for enhanced aesthetics comprehension [31].

Additionally, we present a method to improve automatic image cropping by leveraging aesthetics, geometric composition, and semantics. Our goal is to provide insights into how these methodologies can collectively enhance the performance of IAA tasks, highlighting both their potential and challenges.

Related works

Existing IAA models can be categorized into hand-crafted feature-based and deep learning-based approaches.

Hand-crafted Feature-based IAA

Early IAA studies utilized hand-crafted features to describe image aesthetics. For example, in [9], 56 visual features based on standard photographic rules were extracted to distinguish aesthetically pleasing images. Ke *et al.* [20] analyzed perceptual differences in image quality using high-level semantic features and a Bayes classifier for aesthetic decisions. Marchesotti *et al.* [24] employed generic image descriptors like Bag-of-Visual-words (BOV) and Fisher Vector (FV). Despite their explicit physical meanings, hand-crafted features achieved limited success due to their weak representation of the abstract nature of human aesthetic perception.

Deep Learning-based IAA

With advances in deep learning and large-scale IAA databases, research has shifted towards deep learning-based models. Recent models have explored relationships between image themes and visual attributes, such as the theme-aware visual attribute reasoning model by Li et al. [21]. Pan *et al.* [26] utilized a multi-task framework for simultaneous prediction of aesthetic scores and attributes, employing adversarial learning to enhance model performance. Recognizing that binary classification or single scalar scores do not fully capture human aesthetic perception, recent studies have focused on aesthetic distribution prediction. Talebi *et al.* [30] introduced NIMA, predicting aesthetic score distributions using a fully connected layer and optimizing with Earth Mover's Distance (EMD) loss. The MLSP model [16] used transfer learning from ImageNet for aesthetic predictions. She *et al.* [28] proposed HLA-GCN, integrating image layout information with graph convolutional networks. The AMM-Net model [22] modeled relationships between visual and textual modalities for IAA. Celona *et al.* [6] used an ImageNet-pretrained model to predict image composition and style attributes, with aesthetic predictor weights obtained via an attribute-conditioned hypernetwork.

Methods

Aesthetic of images containing faces

The method for assessing the aesthetic quality of images with faces is depicted in Figure 2 and involves face detection, feature extraction, and feature fusion with Genetic Algorithm (GA) based learning. First, the largest face in an image is detected using the RetinaFace detector [10], with a 10% increase in bounding box size to include shoulders. Next, features are extracted: global image features using the DeepBIQ [4] and DeepIA [3] models for perceptual qualities and aesthetic attributes, and facial features using the AFFACT model [13] for detailed facial attributes. The extracted features are concatenated into a single feature vector. GA is employed to select relevant features and optimize a linear model to predict aesthetic quality. The GA uses mixed coding for chromosomes, combining boolean values for feature selection with real values for model weights and bias. The aesthetic quality prediction is computed by combining the selected features with the weights and bias.

The method addresses both classification and regression problems. For classification, it uses Hinge loss to minimize errors. For regression, it employs Smooth-L1 loss, Norm-in-Norm loss, and Ranking Hinge loss for improved prediction performance.

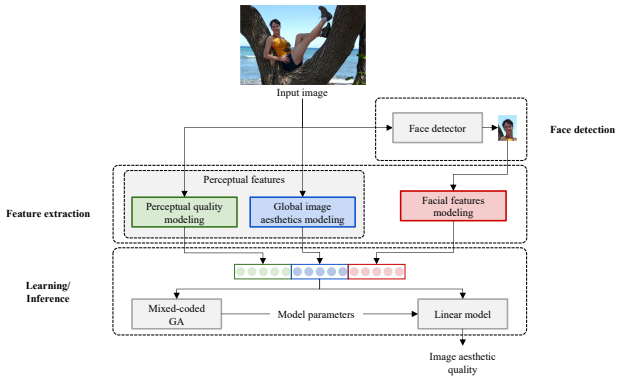

Figure 2: **Overview of the proposed method for portrait image aesthetic assessment [7].** Given an image containing faces, the largest face is detected and cropped. Perceptual features are extracted from the whole image, while facial features are computed on the crop of the face. A mixed-coded Genetic Algorithm (GA) is used for estimating the parameters of a linear model, that predicts the image aesthetic quality.

Aesthetic of generic content images

The semantic content of a photo significantly affects its aesthetic quality. Professional photographers use different techniques and criteria based on the content they are shooting. Similarly, image styles as well as composition rules also influence an image's aesthetic quality. Therefore, we designed a method relying on the categorization of image composition and style for generic content image aesthetic assessment (see Figure 3). Given an input image $\mathbf{X}$, the goal is to estimate both the aesthetic score distribution $\hat{\mathbf{q}}$ and the presence of aesthetic-related attributes $\hat{\mathbf{y}}$ using the network f parameterized with θ^*:

$$\mathbf{X} \xrightarrow{f(\theta^*)} (\hat{\mathbf{q}}, \hat{\mathbf{y}}). \tag{1}$$

The network f consists of two components: f_s and f_t. The network f_s handles the side information regarding aesthetic-related attributes, producing the output $\hat{\mathbf{y}}$ and an embedding $\mathbf{e}_s$. This embedding is used by an attribute-conditioned hypernetwork h to generate the parameters $\hat{\theta}_t$ for the network f_t, which performs the main task of aesthetic assessment.

The parameters θ^* include those learned for the specific tasks and a pre-trained backbone. A two-step optimization procedure introduces attribute constraints into the hypernetwork.

First, the parameters θ_s of the network f_s are trained using a dataset $\mathscr{D}_s$ with images and aesthetic-related attributes. The network f_s predicts whether an attribute occurs in the input image, using an embedding $\mathbf{e}_b$ from a pre-trained backbone.

Second, the hypernetwork is trained to learn the parameters θ_h of a metamodel h, which generates the parameters $\hat{\theta}_t$ for f_t. The training set $\mathscr{D}_t$ contains images and aesthetic score distributions. The goal is to learn θ_h to generate parameters for f_t that predict the aesthetic score distribution $\hat{\mathbf{q}}$.

Proposed Network Architecture

The proposed architecture consists of four networks: the Backbone, the AttributeNet, the HyperNet, and the AestheticNet. The Backbone b encodes the input image $\mathbf{X}$ into a Multi-Level Spatially Pooled (MLSP) embedding vector $\mathbf{e}_b$, using activations from multiple layers of an ImageNet pre-trained neural network. The input image is processed at its original resolution to avoid altering the image composition or aspect ratio, which can harm aesthetics assessment. The embedding vector $\mathbf{e}_b$ is used by both

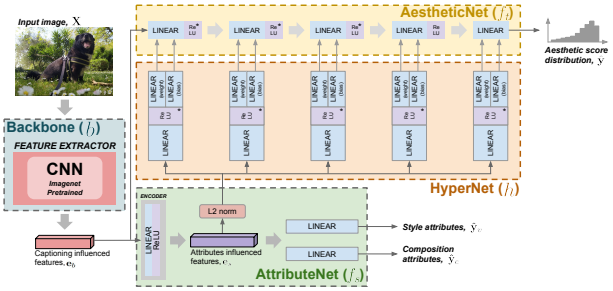

Figure 3: The proposed method for IAA guided by composition and style [6].

the AttributeNet and the AestheticNet. The AttributeNet f_s consists of a Multi-Layer Perceptron (MLP). It is specially trained for mapping the backbone embedding $\mathbf{e}_b$ into K aesthetic-related attributes, namely predictions for image styles ($\hat{\mathbf{y}}_v$) and composition rules ($\hat{\mathbf{y}}_c$). The HyperNet h generates the parameters $\hat{\theta}_t$ for the AestheticNet. Each HyperNet Block (HB) reduces the size of the $l2$-normalized embedding $\mathbf{e}_s$ and estimates the weights and biases for the corresponding layer of the AestheticNet. The AestheticNet f_t predicts the aesthetic score distribution $\hat{\mathbf{q}}$ using the embedding $\mathbf{e}_b$ produced by the Backbone. It is an MLP whose parameters $\hat{\theta}_t$ are computed by the HyperNet.

Multi-modal image aesthetics assessment

Recent datasets [8, 12, 19] extend the IAA problem by including captions related to photo aesthetics and photography skills (see Table 1). These datasets contain N images, each described by K aesthetic critiques c such that $\mathscr{D} = \{(I_1, \mathsf{c}_1^1, \ldots, \mathsf{c}_1^K) \ldots, (I_N, \mathsf{c}_N^1, \ldots, \mathsf{c}_N^K)\}$. Commonly critiqued aspects include composition, subject, use of camera, and color. Novel algorithms now generate aesthetic-oriented critiques for images. These methods map an input image I_i to an aesthetic critique c_k. Although photo critiques provide explicit feedback on aesthetic qualities, their use for classification or image ranking has not been fully explored. Critique generation models face challenges due to the subjective nature of aesthetics, often expressing preferences rather than detailed critiques.

We proposed using sentiment polarity analysis on critiques to define an aesthetic score s_i for images. Given an image I_i and K associated critiques, a sentiment polarity model maps each critique c_k to a vector $\mathbf{p} \in \mathbb{R}^3$, representing probabilities of expressing negative, neutral, or positive sentiment. The sentiment score s_k of a critique is:

$$s_k = \frac{\sum_{l=0}^{2} p_l l}{2},$$

where l represents sentiment labels and p_l the associated probabilities. The overall sentiment score s_i for an image is the average sentiment score of its critiques. The dataset can then be defined as $\mathscr{D} = \{(I_1, \mathsf{c}_1^1, \ldots, \mathsf{c}_1^K, s_1) \ldots, (I_N, \mathsf{c}_N^1, \ldots, \mathsf{c}_N^K, s_N)\}$.

This approach is the first to estimate aesthetic quality directly from critiques rather than ratings. Critiques provide valuable human judgment insights, offering explanations for aesthetic preferences. Our method captures these insights into a quantifiable score, linking Aesthetic Image Captioning (AIC) with IAA. This integration enhances explainability of aesthetic scores and potentially predicts valence-sensitive critiques. Additionally, our weakly-supervised labeling approach relies only on comments, reducing the need for intensive human effort.

We exploit sentiment polarity analysis on aesthetic critiques to define sentiment scores, employing TwitterRoBERTa [2], a deep learning model inspired by RoBERTa [23]. Although trained on Twitter data, the model's performance in social media contexts makes it suitable for our task. Future work may explore models tailored to specific sub-domains. Transformer models for sentiment analysis have biases, warranting deeper analysis of these biases in our application.

We use the TwitterRoBERTa implementation from HuggingFace transformers library [32]. Figure 4 shows samples of our dataset annotated with sentiment scores.

We assess the correlation between sentiment scores and human aesthetic judgments for AVA and PCCD images. Spearman's Rank Correlation Coefficient (SRCC) and Pearson's Linear Correlation Coefficient (PLCC) show positive correlations, indicating the effectiveness of our sentiment score as an approximation of aesthetic quality.

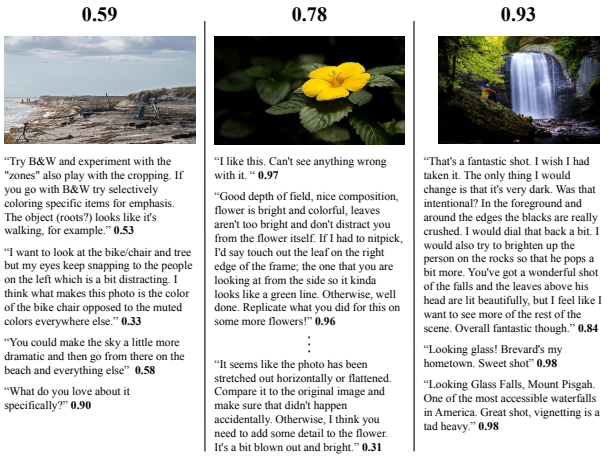

Figure 4: RPCD samples annotated with the proposed sentiment score from [31].

Table 1: Comparison of the properties in different benchmark datasets on image aesthetic captioning.

Dataset	AVA-Comments [35]	DPC-Captions [19]	PCCD [8]	RPCD [31]
Images	253,961	117,132	4,235	73,965
Avg image resolution	607×537	606×534	1414×1202	2993×2716
Attributes	–	5	7	7*
Comments	3,601,761	208,926	29,645	219,790
Comments per image	14.1	1.8	6.6	2.9
Avg words per comment	14.6	24.5	41.1	49.1
Max words per comment	2146	549	780	1286
Content category	66	66	27	6
Rating scale	1-10	1-10	1-10	0-1
Avg raters per image	6	15	7	–

Application

In this section we present an autocropping method that aims to identify a sub-region of the original image that is best in terms of aesthetics, geometric composition, and semantic content preservation. The method, inspired by anchor-based approaches, extends the idea by considering multiple cropping criteria instead of just one. Given an input image and a list of pre-defined anchor boxes, the method uses three strategies based on aesthetics, composition, and semantics to generate three ranked lists of anchor boxes. Each strategy employs a Deep Convolutional Neural Network (CNN) trained for its specific task. The final ranked list of anchor boxes is the average of the three lists, with the best crop being the one with the highest score.

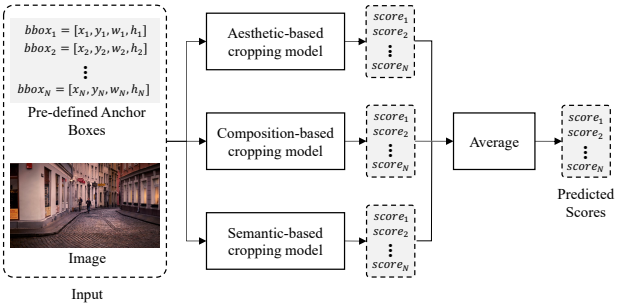

Figure 5: **Overview of the proposed cropping method [5].** Given an image and a set of pre-defined anchor boxes as input, the method estimates a score for each box of the set by averaging the predictions of three models which models aesthetics, geometric composition, and semantics of the crops. The final scores are then ranked to select the best crop.

The Cropping Model

The cropping model involves a lightweight backbone based on the MobileNet-v2 architecture, which extracts multi-scale features from an input image. These features are then processed by a cropping module that uses RoIAlign and RoDAlign operations to extract regions corresponding to each anchor box, producing a score for each box. Multi-scale features are extracted from the CNN to ensure invariance to object scales in photographs. Three different scales are used, with feature volumes concatenated and resized to a uniform spatial resolution. The number of channels is then reduced for efficiency. For each anchor box, the cropping module extracts two regions: the Region of Interest (RoI) and the Region of Discard (RoD). RoIAlign and RoDAlign operations are used to process these regions, which are then fed into a series of convolutional blocks to produce a score. The cropping model is trained using a Smooth-L1 loss.

Cropping Models Ensemble

Many deep learning-based cropping methods consider only a single criterion. However, human preferences are influenced by various factors. We employ three criteria: image aesthetics, composition, and semantics. Each criterion is modeled by pre-training the backbone architecture for its respective task.

The aesthetics model is trained on the AVA dataset using the approach proposed by [30]. The composition model is trained on the KU-PCP dataset to predict geometric composition classes. The semantics model is trained on the ImageNet dataset for image categorization.

Each pre-trained backbone is used to train three cropping models: aesthetics-based, composition-based, and semantics-based. Each model predicts a list of scores for pre-defined anchor boxes. The overall score for each box is the average of the scores from the three models. The final scores are then ranked to select the best crop.

Given an image, the three cropping models predict scores for N anchor boxes, denoted as $\mathbf{s^A}$, $\mathbf{s^C}$, and $\mathbf{s^S}$. The overall score for a box is:

$$s_i^{ASM} = \frac{s_i^A + s_i^C + s_i^S}{3}.$$

The final scores are ranked to select the best crop for the image.

Open Challenges and Future Directions

Explainable Aesthetic Assessment

One of the foremost challenges in the field of IAA is the development of systems that provide explanations for their decisions, moving beyond the typical "black box" approach. The integration of explainable artificial intelligence (XAI) techniques is essential for generating insights into how different features influence aesthetic judgments. Future research should aim to develop methods that elucidate these influences through textual or visual explanations, highlighting the key aspects of images that affect their aesthetic scores. Recent literature addresses this goal by leveraging Multimodal Large Language Models (MLLMs), which, by processing and generating linguistic, visual, and other modality data, can provide richer and more context-aware aesthetic evaluations [33, 36].

Multi-modal Aesthetic Assessment

As digital content increasingly becomes multi-modal, incorporating text, audio, and interactive elements alongside visual components, there is a need to extend IAA to these complex formats. Multi-modal aesthetic assessment could offer a more comprehensive evaluation of multimedia content, such as videos with audio tracks or interactive web pages. This challenge necessitates the creation of innovative frameworks and models capable of integrating diverse sensory information to assess overall aesthetic appeal comprehensively.

New Applications of Aesthetic Assessment

The scope of IAA applications is expanding into areas like real-time video content curation for streaming platforms, design of virtual and augmented reality environments, and integration with robotic systems for tasks requiring aesthetic sensitivity, such as interior design. These new applications present opportunities for IAA to influence user engagement significantly, enhance user experience in immersive environments, and improve functionality in autonomous service systems.

Conclusion

This paper has synthesized findings from three innovative approaches to IAA, providing a comprehensive overview of the field's current state. Our discussions included the use of genetic algorithms for feature combination in images with faces, the application of composition and style attributes for aesthetic prediction, and the integration of NLP with computer vision to enhance aesthetic understanding. Furthermore, we introduced a method that leverages aesthetics, geometric composition, and semantics to improve automatic image cropping. These methodologies collectively aim to enhance the performance of IAA tasks, addressing both their potential and associated challenges.

Moreover, recent advancements in explainability, multi-modal analysis, and the exploration of new applications promise to significantly enhance the capabilities of IAA systems. These developments are poised to expand the practical utility of IAA across various industries, making it an essential tool in the design of engaging and aesthetically pleasing digital media and environments.

References

[1] Mirko Agarla, Luigi Celona, Raimondo Schettini, et al. Predicting video memorability using a model pretrained with natural language supervision. In *MediaEval Multimedia Benchmark Workshop Working Notes*, volume 1. CEUR Workshop Proceedings, 2023.

[2] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650. Association for Computational Linguistics, 2020.

[3] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. Predicting image aesthetics with deep learning. In *International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS)*, pages 117–125. Springer, 2016.

[4] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. On the use of deep learning for blind image quality assessment. *Springer Signal, Image and Video Processing*, 12(2):355–362, 2018.

[5] Luigi Celona, Gianluigi Ciocca, and Paolo Napoletano. A grid anchor based cropping approach exploiting image aesthetics, geometric composition, and semantics. *Expert Systems with Applications*, 186:115852, 2021.

[6] Luigi Celona, Marco Leonardi, Paolo Napoletano, and Alessandro Rozza. Composition and style guided image aesthetic assessment. *IEEE Transactions on Image Processing*, 2022.

[7] Luigi Celona and Raimondo Schettini. A genetic algorithm to combine deep features for the aesthetic assessment of images containing faces. *Sensors*, 21(4):1307, 2021.

[8] Kuang-Yu Chang, Kung-Hung Lu, and Chu-Song Chen. Aesthetic critiques generation for photos. In *ICCV*, pages 3514–3523. IEEE, 2017.

[9] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, pages 288–301. Springer, 2006.

[10] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, pages 5203–5212. IEEE, 2020.

[11] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Aesthetic-driven image enhancement by adversarial learning. In *International Conference on Multimedia*, pages 870–878, 2018.

[12] Koustav Ghosal, Aakanksha Rana, and Aljosa Smolic. Aesthetic image captioning from weakly-labelled photographs. In *ICCV Workshops*, pages 0–0. IEEE/CVF, 2019.

[13] Manuel Günther, Andras Rozsa, and Terrance E. Boult. Affact - alignment free facial attribute classification technique. In *International Joint Conference on Biometrics (IJCB)*, 2017.

[14] Guanjun Guo, Hanzi Wang, Chunhua Shen, Yan Yan, and Hong-Yuan Mark Liao. Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression. *IEEE Transactions on Multimedia*, 20(8):2073–2085, 2018.

[15] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. The interestingness of images. In *ICCV*, pages 1633–1640. IEEE, 2013.

[16] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *CVPR*, pages 9375–9383. IEEE/CVF, 2019.

[17] Jin Huang, Yongshun Gong, Lu Zhang, Jian Zhang, Liqiang Nie, and Yilong Yin. Modeling multiple aesthetic views for series photo selection. *IEEE Transactions on Multimedia*, 2023.

[18] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *CVPR*, pages 145–152. IEEE, 2011.

[19] Xin Jin, Le Wu, Geng Zhao, Xiaodong Li, Xiaokun Zhang, Shiming Ge, Dongqing Zou, Bin Zhou, and Xinghui Zhou. Aesthetic attributes assessment of images. In *International Conference on Multimedia*, pages 311–319. ACM, 2019.

[20] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *CVPR*, volume 1, pages 419–426. IEEE, 2006.

[21] 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.

[22] Leida Li, Tong Zhu, Pengfei Chen, Yuzhe Yang, Yaqian Li, and Weisi Lin. Image aesthetics assessment with attribute-assisted multimodal memory network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[24] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *2011 international conference on computer vision*, pages 1784–1791. IEEE, 2011.

[25] L Neumann, M Sbert, B Gooch, W Purgathofer, et al. Defining computational aesthetics. *Computational aesthetics in graphics, visualization and imaging*, pages 13–18, 2005.

[26] Bowen Pan, Shangfei Wang, and Qisheng Jiang. Image aesthetic assessment assisted by attributes through adversarial learning. In *AAAI*, volume 33, pages 679–686, 2019.

[27] Lauren Scissors, Moira Burke, and Steven Wengrovitz. What's in a like? attitudes and behaviors around receiving likes on facebook. In *Conference on Computer-supported Cooperative Work & Social Computing*, pages 1501–1510. ACM, 2016.

[28] Dongyu She, Yu-Kun Lai, Gaoxiong Yi, and Kun Xu. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In *CVPR*, pages 8475–8484. IEEE/CVF, 2021.

[29] Wei-Tse Sun, Ting-Hsuan Chao, Yin-Hsi Kuo, and Winston H Hsu. Photo filter recommendation by category-aware aesthetic learning. *IEEE Transactions on Multimedia*, 19(8):1870–1880, 2017.

[30] H. Talebi and P. Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

[31] Daniel Vera Nieto, Luigi Celona, and Clara Fernandez Labrador. Understanding aesthetics with language: A photo critique dataset for aesthetic assessment. In *Advances in Neural Information Processing Systems*, volume 35, pages 34148–34161. Curran Associates, Inc., 2022.

[32] Thomas Wolf and etal. Transformers: State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, Oct. 2020.

[33] Zhiwei Xiong, Yunfan Zhang, Zhiqi Shen, Peiran Ren, and Han Yu. Multi-modal learnable queries for image aesthetics assessment. In *International Conference on Multimedia and Expo*. IEEE, 2024.

[34] Fang-Lue Zhang, Miao Wang, and Shi-Min Hu. Aesthetic image enhancement by dependence-aware object recomposition. *IEEE Transactions on Multimedia*, 15(7):1480–1490, 2013.

[35] Ye Zhou, Xin Lu, Junping Zhang, and James Z Wang. Joint image and text representation for aesthetics analysis. In *International Conference on Multimedia*, pages 262–266. ACM, 2016.

[36] Zhaokun Zhou, Qiulin Wang, Bin Lin, Yiwei Su, Rui Chen, Xin Tao, Amin Zheng, Li Yuan, Pengfei Wan, and Di Zhang. Uniaa: A unified multi-modal image aesthetic assessment baseline and benchmark. *arXiv preprint arXiv:2404.09619*, 2024.