A Comprehensive Survey on Image Aesthetic Caption

Guanjun Sheng

***Abstract*—With the continuous development of deep learning in recent years, the field of image aesthetic caption has gradually become a popular research direction, which significantly impacts various applications, such as advanced semantic understanding of images and the promotion of artistic images. Image aesthetic caption generation is a cutting-edge direction for integrating computer vision and natural language processing. Unlike traditional image caption that focuses on outputting basic facts of images, the core goal of image aesthetic caption research is to generate image aesthetic description texts that combine semantic accuracy and artistic expression through deep learning algorithms. In response to the problem that there is no review article in this field, this article systematically reviews the technical development context in the field of image aesthetic caption, focusing on analyzing the technical route of image aesthetic caption based on traditional deep learning architecture, namely convolutional neural networks and recurrent neural networks, and also introduces the technical application of multimodal models in the field of image aesthetic caption in recent years. This article covers the main technical methods, data sets, evaluation indicators, and future development trends in the field of image aesthetic caption and analyzes the challenges and opportunities of current research. We hope that our review can be a reference for future research in the field of image aesthetic caption.**

***Index Terms*—Image Aesthetic Caption, Image-to-text, Overview, Deep Learning.**

I. INTRODUCTION

In today's era of rapid development of digital media, social media has become an essential part of people's daily lives. People usually share their pictures on social media, so images have become a crucial carrier in spreading and expressing information in the digital age. Then, how to evaluate the aesthetic value of these images has become a hot research topic. The traditional image caption generation task focuses on the objective content of the image, such as object categories, scene information, and action descriptions, while ignoring the subjective aesthetic characteristics of the image, such as color matching, composition art, emotional expression, and style preferences [1]–[3]. Fig 1 shows the difference between image caption research and image aesthetic caption research. With the rapid development of multimedia technology, image aesthetic caption generation, as the intersection of computer vision and natural language processing, is gradually evolving from functional tools to a medium that combines information transmission and artistic expression. Its core goal is to transform visual content into natural language descriptions through algorithms and, at the same time, impart aesthetic value to text in dimensions such as font design, color matching, and dynamic interaction, thereby improving the accessibility and emotional resonance of the content. This technology not only provides key support for visually impaired people, education, and medical care but also creates a new form of artistic expression in scenes such as film and television communication and social media.

Image Caption: There is a waterfall in the woods

Image Aesthetic Caption : This picture is well composed and the lighting is also very good.

Fig 1. Difference between Image Caption and Image Aesthetic Caption

In recent years, breakthroughs in deep learning technology have promoted paradigm innovation in this field. From early template matching and keyword extraction to end-to-end models based on encoder-decoder architecture to cross-modal alignment methods that integrate attention mechanisms and multimodal pretraining (such as CLIP and Transformer), the technological evolution has significantly improved the semantic accuracy and scene adaptability of captions. The field of image aesthetic caption has been based on deep learning in the early days, using traditional convolutional neural networks (CNNs) to extract image aesthetic features and using recurrent neural networks (RNNs) and long and short-term memory networks (LSTMs) to achieve the generation of image aesthetic caption [4]–[7]. In recent years, with the continuous development of large-model technology, research methods in image aesthetic caption have gradually approached multimodal large-scale models [8]–[13]. This article aims to systematically sort out the technological evolution and aesthetic innovation in the field of image aesthetic caption, also focuses on discussing the relevant data sets and evaluation standards in the field of image aesthetic caption, and further explores the challenges it faces and the development trends in the field of image aesthetic caption in the field of future image aesthetic caption.

Manuscript received April 09, 2025

Guanjun Sheng, School of Computer Science and Technology, Tiangong University, Tianjin, 300387, China

In Chapter 2, we focus on the relevant datasets in the field of image aesthetic caption. In Chapter 3, we systematically review the technological development route in image aesthetic caption, from traditional deep learning to multimodal large models. In Chapter 4, we summarize this article.

II. DATASET

A. PCCD dataset (Photo Critique Captioning Dataset)

The PCCD dataset [1] is the first public dataset focusing on the photo aesthetic comment generation task. It contains 4,235 images and 61,702 reviews provided by professional photographers. Each review corresponds to seven aesthetic dimensions, including overall impression, composition and perspective, color and lighting, theme, depth of field, focus, camera usage, exposure, and shutter speed. Each data point is a triple in the form (picture, comment, aesthetic aspect), and each comment is also equipped with an aesthetic score that has been normalized (from 1 to 10 to [0,1]) to indicate the attractiveness of this aesthetic element. The case of the PCCD data set is shown in Fig 2, for example. The data set breaks through the limitation of traditional aesthetic scores that only judge "good or bad", and deeply analyzes photography skills and aesthetic details through multi-angle text descriptions (such as "the white space on the right needs to be optimized", "the vanishing points and lines are handled well", etc.). The experiment focuses on the three dimensions of high-frequency composition, color, and subject and selects 3,840 images and 30,254 sentences for training. Combined with MSCOCO pre-training to improve the essential description ability of the model, it aims to generate diverse and professional photography improvement suggestions, providing critical data support for the research on aesthetic semantic generation in computer vision. The PCCD dataset provides rich and highly targeted training data for generating image aesthetic comments. It allows binary judgments on the aesthetic quality of the picture and generates specific and detailed aesthetic feedback, which helps improve photography skills and aesthetic level.

B. AVA-Captions dataset

The AVA-Captions dataset [4] is a large-scale aesthetic image description dataset built by cleaning and optimizing user comments in AVA (Aesthetic Visual Analysis) raw data. The original AVA contains about 250,000 photography community pictures and 3 billion comments, but these comments have misspellings, grammatical problems, and a lot of low-information content (such as "Great Photos") [14]. To solve this problem, the researchers proposed a probability n-gram filtering strategy [15], calculating the corpus frequency of single-word (noun) and double-word ("descriptive word-object" combination), combining the negative log probability formula to score the comments, filtering out 55% of low-quality content, and finally retaining about 230,000 pictures and 1.15 million high-quality comments (on average 5 per figure). The scale of this dataset is 60 times that of PCCD, the only aesthetic description dataset which covers diverse aesthetic attributes such as composition, color, and post-processing. It also verifies its information volume and accuracy through subjective evaluation. It can be used as a benchmark dataset for tasks such as aesthetic image analysis and intelligent photography equipment development.

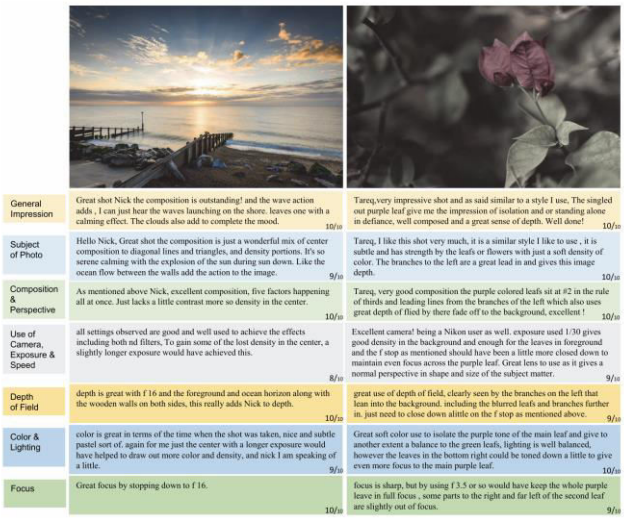

Fig 2. Sample in the PCCD dataset [1].

C. DPC-Captions dataset

The DPC-Captions dataset [5] is a large-scale weak label dataset built for the multi-attribute evaluation task of image aesthetics. The data comes from 330,000 images and corresponding comments crawled by the DPChallenge.com website. Aesthetic keywords are extracted from the fully labeled small-scale PCCD dataset (including 7 aesthetic attribute scores and comments) for labeling and filtering through knowledge migration. The dataset finally contains 154,384 images and 2,427,483 comments, and forms 5 core aesthetic attributes by merging relevant attributes (such as "Depth of Field" and "Focus", "General Impression" and "Subject of Photo"): color and lighting, composition, depth of field and focus, impression and theme, and camera use. Comments for each image are classified into specific attributes through keyword matching (such as "composition" corresponding to composition attributes), ensuring that the content of the comment is highly relevant to the attributes. Compared with other aesthetic datasets (such as AVA-Reviews that only contain single comments and AVA-Comments that do not have attributes), DPC-Captions are not only larger in scale (on average 15 comments per figure), but also implement multi-attribute fine-grained annotation for the first time. The data set is divided into training/verification/test sets by attributes. Each attribute reserves 2,000 samples for verification and testing, providing a data basis for joint learning of image aesthetic attribute description and scoring.

D. DPC2022 dataset

DPC2022 is a large-scale multimodal dataset [8] designed for Image Aesthetic Quality Assessment (AQA) and Image Aesthetic Caption Generation (IAC) studies, containing 510,000 images, more than 5 million user reviews, and 350,000 aesthetic ratings of 1-10 points. This dataset is constructed by crawling and cleaning the data accumulated by the professional photography community dpchallenge.com over ten years. It uses the spaCy tool to eliminate emojis, abnormal spellings, and redundant symbols,

and ultimately retains high-quality images and normalized text. The data set is divided into two subsets: SetA contains all 510K images and comments for IAC tasks; SetB contains 350K images with both comments and scores for multimodal AQA research. Statistics show that each image has an average of 10 comments, each comment contains about 21 sentences, and the average sentence length is 19 words. As the largest graphic and text aesthetic database at present, DPC2022 significantly surpasses traditional data sets such as AVA and PCCD. Its rich visual-text pairs and fine-grained scoring systems provide a new benchmark for cross-modal aesthetic calculations. This dataset has been open source and has shown significant value in multimodal AQA baseline testing, aesthetic text correlation verification (ARS) and ARIC model training, promoting the joint research of aesthetic text generation and score prediction.

III. Method

In this section, we will review the development of image aesthetic caption from the perspectives of research based on traditional deep learning and research based on large models. Most research on image aesthetic caption based on traditional deep learning uses CNN to extract aesthetic features from images and uses RNN or LSTM to generate image aesthetic captions. Research on image aesthetic caption based on large models has a cross-modal feature extraction step for images.

A. Research on image aesthetic caption based on traditional deep learning

In 2017, Chang et al. [1] proposed a novel method for generating photo aesthetic reviews, breaking through the limitations of traditional binary evaluation of aesthetic quality and achieving the task of generating professional photography reviews from multiple dimensions for the first time. The researchers developed two deep learning models: the AO (Aspect-Oriented) method based on a divide-and-conquer strategy trains CNN-LSTM models for specific aesthetic dimensions such as composition and color. In contrast, the AF (Aspect-Fusion) method dynamically fuses hidden features of different aesthetic dimensions through an attention mechanism to generate richer and more coherent reviews. To support this research, the team constructed the first public professional photography review dataset, PCCD, which contains 4235 images and more than 60,000 multi-dimensional reviews. Experiments show that the AF method outperforms traditional methods in both the automatic evaluation index SPICE and manual evaluation. The generated comments not only have a 28% increase in semantic accuracy but also a 66% increase in diversity, and can effectively provide specific improvement suggestions. This work combines computer vision with natural language processing, opening up new directions for photography education and technical analysis.

In 2019, Ghosal et al. [4] proposed an innovative method for generating aesthetic image caption (AIC) from weakly annotated online photos. In response to the problem of lack of large-scale annotated data in existing aesthetic description tasks, the authors developed an automatic cleaning strategy based on probabilistic n-gram filtering [15], and constructed the AVA-Captions dataset (containing 230,000 images and 1.5 million cleaned descriptions) from the AVA dataset (containing 250,000 photos and 3 billion user comments). The data quality was significantly improved by retaining the "descriptor-object" combination sentences with rich information. At the same time, they proposed a weakly supervised CNN training method, using latent Dirichlet allocation (LDA) to automatically discover 200 potential aesthetic themes (such as "motion blur", "black and white contrast", etc.) from the comments, replacing the traditional ImageNet [16] supervised training. Experiments show that this method outperforms the noisy data benchmark in terms of BLEU, CIDEr, and other indicators, generates more diverse and accurate descriptions, and shows good generalization ability in cross-dataset (PCCD) tests. Subjective evaluation also verifies the consistency of the cleaning strategy with human judgment, providing a new benchmark dataset and weakly supervised training paradigm for aesthetic analysis tasks.

In the same year, Jin et al. [5] proposed a new image aesthetic evaluation method, which achieves a more comprehensive image aesthetic analysis by simultaneously generating multi-attribute aesthetic descriptions and scores. The authors constructed a large-scale weakly annotated dataset DPC-Captions (including 154,000 images and 2.42 million comments), and extracted five types of aesthetic attributes (such as composition, color and lighting, etc.) from the comments of DPChallenge.com through a keyword migration strategy, making up for the small scale of the existing dataset PCCD and the lack of attribute annotations in AVA. The proposed multi-attribute aesthetic network (AMAN) adopts a two-stage training strategy, combining multi-task feature extraction, channel-space attention mechanism and LSTM language generation, and integrates fully annotated PCCD data (including attribute scores) and weakly annotated DPC-Captions data in feature learning. Experiments show that AMAN outperforms traditional CNN-LSTM and SCA-CNN models in both image description generation tasks (indicators such as BLEU, METEOR) and attribute scoring tasks (mean square error), and achieves fine-grained aesthetic attribute analysis and interpretability evaluation.

In 2020, Xiong et al. [6] proposed a personalized aesthetic image caption generation method (PAIC) to solve the problems of missing user preferences and insufficient feature expression in existing aesthetic image caption (AIC) technologies. Traditional AIC methods mainly rely on high-level semantic features of images, ignore the impact of low-level visual features (such as color and composition) on aesthetic style, and do not consider differences in user subjective evaluation. To this end, PAIC designs three core modules: 1) Aesthetic Feature Extraction Network (AEN), which fuses multi-level visual features through multi-level spatial pooling (MLSP) and multi-column CNN; 2) User Encoding Network (UEN), which uses TF-IDF and LSTM to extract explicit vocabulary preferences and implicit embedding vectors from user historical comments; 3) Personalized Description Model, which combines user preference vectors with visual features and generates personalized descriptions by dynamically adjusting vocabulary selection and visual attention. The experiment was based on the newly constructed AVA-PCap dataset (containing 450,000 user-image-comment data). The results

showed that PAIC outperformed the baseline model by more than 10% in terms of BLEU, CIDEr and other indicators, verifying the effectiveness of user preference modeling and multi-level feature fusion. This study integrated personalization into the AIC task for the first time, providing a solution to the diverse aesthetic needs in practical applications.

In 2021, Yeo et al. [7] proposed a deep learning-based framework for generating aesthetic reviews of photographic works. Unlike traditional image descriptions (which focus on objective content), this method captures the content and aesthetic attributes of images by integrating two feature encoders, namely, a fact feature encoder (ResNet [17]) based on image classification tasks and an aesthetic feature encoder (trained using Earth Mover's Distance loss [18]) based on aesthetic scoring tasks. Comments are generated by combining multi-encoder fusion and attention mechanisms with an LSTM decoder. The experiment was verified on the AVA-Captions dataset, and the results showed that the method outperformed the baseline model in terms of METEOR [19] and ROUGE indicators, especially in terms of semantic diversity (assessed by Word Mover's Distance) and synonym usage, indicating that it can generate diversified comments that are closer to human subjective evaluations. The study also demonstrated the limitations of traditional n-gram indicators (such as BLEU) for aesthetic review tasks, and proposed WMD as an alternative that is more in line with semantic similarity evaluation.

B. Research on image aesthetic caption based on large models

In 2022, Zhong et al. [8] proposed an aesthetically relevant image caption generation method (ARIC), combining image aesthetic quality assessment (AQA) and image aesthetic caption (IAC) to solve the problem of insufficient aesthetic relevance between text description and existing methods. The authors first introduced the aesthetic relevance score (ARS), quantified the aesthetic value of text through five dimensions: aesthetic vocabulary, sentence length, object vocabulary, sentiment score, and TF-IDF, and automatically predicted ARS based on the BERT model [20]. Based on ARS, a weighted loss function and a diversified caption selector (DACS) were designed to optimize the model to generate more aesthetically relevant and diverse descriptions. At the same time, a DPC2022 dataset containing 510,000 images, 5 million comments, and 350,000 aesthetic ratings was constructed. Experiments show that high ARS text can more accurately predict aesthetic ratings, and ARIC outperforms traditional methods in generation accuracy, aesthetic relevance, and diversity, verifying the effectiveness of ARS and the potential of multimodal AQA.

In the same year, Ke et al. [9] proposed a visual-language aesthetic learning framework called VILA, which learns image aesthetics through user comments rather than manual ratings. The method is divided into two stages: first, multimodal pretraining is performed on image-comment pairs (LAION-5B [21] and AVA-Captions) based on the CoCa architecture [22], combining contrastive learning and generation objectives to enable the model to capture rich aesthetic semantics such as composition, color, and style; then, a lightweight ranking-based adapter [23] is designed (only 0.1% of parameters are adjusted), and the image embedding space is adjusted through text anchors (such as "good images") to convert aesthetic ratings into a ranking problem. Experiments show that VILA performs best in the aesthetic description task of AVA-Captions, and its zero-shot capability surpasses supervised models on AVA-Style style classification (69% mAP) and AVA dataset IAA (SRCC 0.657). After fine-tuning with the adapter, the model achieved SOTA (SRCC 0.774) on the AVA dataset, verifying the effectiveness of learning aesthetic representations from unlabeled reviews and significantly reducing the data annotation cost.

IV. Conclusion

This paper systematically reviews the technical development and research progress in the field of image aesthetic caption, and introduces the evolution of the technical route in this field from traditional deep learning models to multimodal large models. The article sorts out the breakthroughs of core datasets such as PCCD and AVA-Captions in multi-attribute annotation and noise processing, and analyzes the limitations of traditional evaluation indicators such as BLEU and CIDEr in measuring the diversity of aesthetic expression. It points out the challenges that current research still faces, such as strong subjectivity in annotation, imperfect evaluation system, and insufficient lightweight model. It further proposes the need to build a professional annotation system, develop aesthetic perception evaluation indicators, and explore efficient model architectures to promote the application of this technology in photography education, smart devices and other scenarios. It is hoped that the research in this article will be of some help to future research in the field of image aesthetic caption.

References

[1] Chang K Y, Lu K H, Chen C S. Aesthetic critiques generation for pho-tos[C]//Proceedings of the IEEE international conference on computer vision. 2017: 3514-3523.

[2] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//International conference on machine learning. PMLR, 2015: 2048-2057.

[3] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption gen-erator[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3156-3164.

[4] Ghosal K, Rana A, Smolic A. Aesthetic image captioning from weakly-labelled photographs[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019: 0-0.

[5] Jin X, Wu L, Zhao G, et al. Aesthetic attributes assessment of imag-es[C]//Proceedings of the 27th ACM international conference on multimedia. 2019: 311-319.

[6] Xiong K, Jiang L, Dang X, et al. Towards personalized aesthetic image cap-tion[C]//2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020: 1-8.

[7] Yeo Y Y, See J, Wong L K, et al. Generating aesthetic based critique for photographs[C]//2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021: 2523-2527.

[8] Zhong Z, Zhou F, Qiu G. Aesthetically relevant image caption-ing[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(3): 3733-3741.

[9] Ke J, Ye K, Yu J, et al. Vila: Learning image aesthetics from user comments with vision-language pretraining[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 10041-10051.

[10] Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International confer-ence on machine learning. PMLR, 2022: 12888-12900.

[11] Chen J, Guo H, Yi K, et al. Visualgpt: Data-efficient adaptation of pretrained lan-guage models for image captioning[C]//Proceedings of the IEEE/CVF Confer-ence on Computer Vision and Pattern Recognition. 2022: 18030-18040.
[12] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International confer-ence on machine learning. PMLR, 2023: 19730-19742.
[13] Zhu D, Chen J, Shen X, et al. Minigpt-4: Enhancing vision-language understand-ing with advanced large language models[J]. arXiv preprint arXiv:2304.10592, 2023.
[14] Murray N, Marchesotti L, Perronnin F. AVA: A large-scale database for aesthetic visual analysis[C]//2012 IEEE conference on computer vision and pattern recog-nition. IEEE, 2012: 2408-2415.
[15] Brown P F, Della Pietra V J, Desouza P V, et al. Class-based n-gram models of natural language[J]. Computational linguistics, 1992, 18(4): 467-480.
[16] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.
[17] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[18] Rubner Y, Tomasi C, Guibas L J. The earth mover's distance as a metric for image retrieval[J]. International journal of computer vision, 2000, 40: 99-121.
[19] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005: 65-72.
[20] Devlin J, Chang M W, Lee K, et al. Bert: Pretraining of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.
[21] Schuhmann C, Beaumont R, Vencu R, et al. Laion-5b: An open large-scale da-taset for training next generation image-text models[J]. Advances in Neural In-formation Processing Systems, 2022, 35: 25278-25294.
[22] Yu J, Wang Z, Vasudevan V, et al. Coca: Contrastive captioners are image-text foundation models[J]. arXiv preprint arXiv:2205.01917, 2022.
[23] Wu Z, Xiong Y, Yu S X, et al. Unsupervised feature learning via non-parametric instance discrimination[C]//Proceedings of the IEEE conference on computer vi-sion and pattern recognition. 2018: 3733-3742.