IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 2, 2021, accepted May 17, 2021, date of publication May 24, 2021, date of current version June 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3083075

A Comprehensive Survey on Computational Aesthetic Evaluation of Visual Art Images: Metrics and Challenges

JIAJING ZHANG, YONGWEI MIAO, (Member, IEEE), AND JINHUI YU
Department of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China
State Key Laboratory of Computer-aided Design and Computer Graphics, Zhejiang University, Hangzhou 310058, China

Corresponding author: Jiajing Zhang (zhangjj@zstu.edu.cn)

This work was supported in part by the Natural Science Foundation of Zhejiang Province under Grant LQ20F020022, and in part by the National Natural Science Foundation of China under Grant 61772463 and Grant 61972458.

ABSTRACT Computational image aesthetic evaluation is a computable human aesthetic perception and judgment realized by machines, which has a significant impact on a variety of applications such as image advanced search and promotional exhibition of painting arts. Various approaches have been proposed in copious literature trying to solve this challenging problem. However, there have been few attempts in reviewing works from different types of visual arts, due to their significant differences in visual features and aesthetic principles. In this survey, we present a comprehensive listing of the reviewed works on aesthetic assessment of photographs and paintings, mainly highlighting the contributions and innovations of the existing approaches. We firstly introduce aesthetic assessment benchmark datasets in different categories. Then, conventional aesthetic evaluation approaches based on handcrafted features are reviewed. Besides, we systematically evaluate recent deep learning techniques that are useful for developing robust models for aesthetic prediction tasks in scoring, distribution, attribute, and description. Moreover, the possibility of aesthetic-aware color enhancement, recomposition of photo images, and automatic generation of aesthetic-guided art paintings through computational approaches are summarized. Finally, challenges and potential future directions for this field are discussed. We hope that our survey could serve as a comprehensive reference source for future research on computational aesthetics in visual media.

INDEX TERMS Computational aesthetic evaluation, visual art image, handcrafted features, deep neural networks, aesthetic enhancement.

I. INTRODUCTION

Aesthetics is an important discipline in visual arts that research on the aesthetic categories such as beauty and ugliness, human aesthetic consciousness, aesthetic experience, creation, development and law of beauty [1]. As an increasing number of visual artworks are created, stored and propagated online in digital forms, the efficient, automated, and quantitative aesthetic evaluation capability has a profound impact on the applications of advanced image retrieving, photo aesthetic enhancement, promotional exhibition of digital painting gallery, and computer-aided creation of art paintings. For instance, when a user enters ''natural lanscape'', he/she will hope to see colorful, pleasing sky and grass views or well-captured mountain instead of gray or blurry snapshots; common people now have more opportunities to appreciate art works casually on online library like Google Art Project without going to museums, it is valuable in helping the amateurish users to better appreciate and understand art beauties.

As one of the natural attributes in human perception, there exist some inspirations from psychological and neuroscience to modern visual aesthetics quantitatively [2]. Researchers find out the diversity of aesthetic simulations between different regions of the brain when people appreciating artworks, and there is a connection between human aesthetic experience and the feeling caused by visual stimuli, which is evoked by activations in distinct and specialized areas of the visual cortex [3]–[6]. These activations can be classified into the early processing of visual stimuli inputs including color, shape, line, orientation, spatial position, and movement,

The associate editor coordinating the review of this manuscript and approving it for publication was Shiqi Wang.

intermediate element combination like grouping and categorization, and late aesthetic perceptual stimulus [7], [8]. Photographers or artists intentionally combine such attributes to form a set of well-established photographic rules, to capture high-fidelity and attractive images, or art principles when they create artworks in order to induce pleasing or desired emotional effects to a large group of audiences [9], [10], as shown in Figure 1.

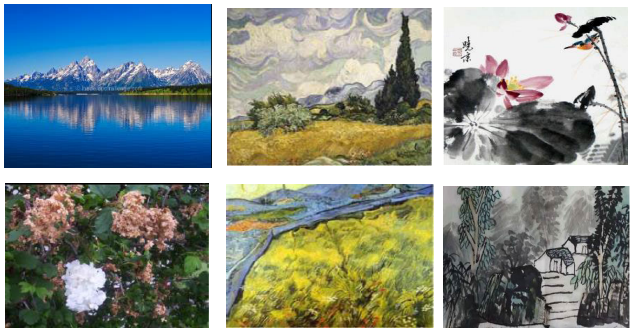


FIGURE 1. Visual art image with different aesthetic qualities. Left: photograph images. Middle: western oil paintings. Right: Chinese ink paintings. The high and low aesthetic quality images are on the top and bottom rows, respectively.

Leder *et al.* [11] proposed a five stages multi-level information processing model of aesthetic experience: perception of low-level information including color, contrast, and complexity, implicit integration of personal experience and memory, explicit classification, cognitive mastering and evaluation, with ultimately aesthetic judgment and aesthetic emotion as production. However, it is difficult to model this sequence computationally for visual art images. The challenges include (i) quantitative modeling the complicated photographic rules, or abstract art principles and appreciation languages, (ii) explaining the aesthetical differences in various images contents, subjects, or art genres (e.g. animal, still life, scenery, architecture, landscape, portrait), (iii) knowing the expression techniques used in capturing photos or drawing paintings (e.g. lighting, sharpness, depth-of-field, motion blur, colorfulness, ink shading, whitespace), and (iv) building a large scale of aesthetic evaluation datasets for various types of visual art images.

To address these challenges, the concept of Computational Aesthetics is proposed at the 1st Eurographics Workshop on Computational Aesthetics in Graphics, Visualization and Imaging in 2005, which aims to use the computing power of the computer itself to replace the processing and analysis process of human brain and visual perception system, to understand various attributes of art images and aesthetic cognition objectively and quantitatively, and simulate the human visual system and perception to make applicable aesthetic judgment on images automatically [12], [13]. It has recently attracted a lot of interest and has become an active research direction in the computer vision field. Previous works mostly researched on photographs and paintings, based on the acquisition of various image aesthetic assessment datasets, early studies [14]–[17] typically adopted rule-based approaches and traditional machine learning algorithms to extract handcrafted low-level features, thus explicitly modeling the photographic techniques or painting principles. However, due to the vagueness of certain photographic or art rules, the handcrafted features are often difficult in approximating computationally. Beginning with the strong performance of Krizhevsky *et al.* [18] in the image classification, the growing amount of datasets, and feasible transfer learning with fine-tuned networks [19], deep learning methods [20], [21] have been applied to aesthetic quality assessment of visual art images, which can automatically learn effective aesthetic features with successful attempts and promising results [22]–[25].

Recently, there are some surveys about computational image aesthetic analysis. Joshi *et al.* [26] discussed key aspects of computational inference of aesthetics and emotion from images, but the deep learning based methods are not considered. Some reviews focused on image quality assessment [27]–[29], however, the image quality usually relates to distortions caused by lossy compression, noises, transmission channel attenuation, and distinguishes noisy images from clean images in terms of different quality measures such as structural similarity index [30], visual signal-to-noise ratio [31], and visual information fidelity [32], rather than photographic or artistic aesthetics. Deng *et al.* [33] systematically reviewed approaches based on visual feature types, dataset characteristics, evaluation metrics, and also conducted experiments to compare the predictive performances of various deep learning settings. But it do not relate to different aesthetic tasks, especially aesthetic distribution, aesthetic factors, aesthetic description. Besides, it mainly focuses on photographic images without considering art-related painting images. Bai *et al.* [34] summarized research methods and evaluation indicators in experimental aesthetics, associated emotion, complexity, artist, and style classification in computational aesthetics of painting images. Fiorucci *et al.* [35] reviewed the research methods of machine learning in painting attribute recognition, forgery identification, and art history. From the perspective of physics and math, Perc [36] analyzed the quantitative evaluation of beauty in culinary art, painting art, and music art. Lu *et al.* [37] summarized the computational aesthetics of fine art paintings into attribute recognition, content understanding, and aesthetic judgments according to the key processes of human aesthetics that include perception, cognition, and evaluation. However, these review articles only discussed approaches in art paintings without covering natural photo images.

Due to the significant differences in visual features and aesthetic principles of different visual arts, in this survey, we would like to contribute an extensive listing of the reviewed works on aesthetic assessment of photographs and paintings, mainly comparing the contributions and innovations of the existing approaches between these two typical types of visual arts, and discuss challenges and potential insights for future directions in this field of study.

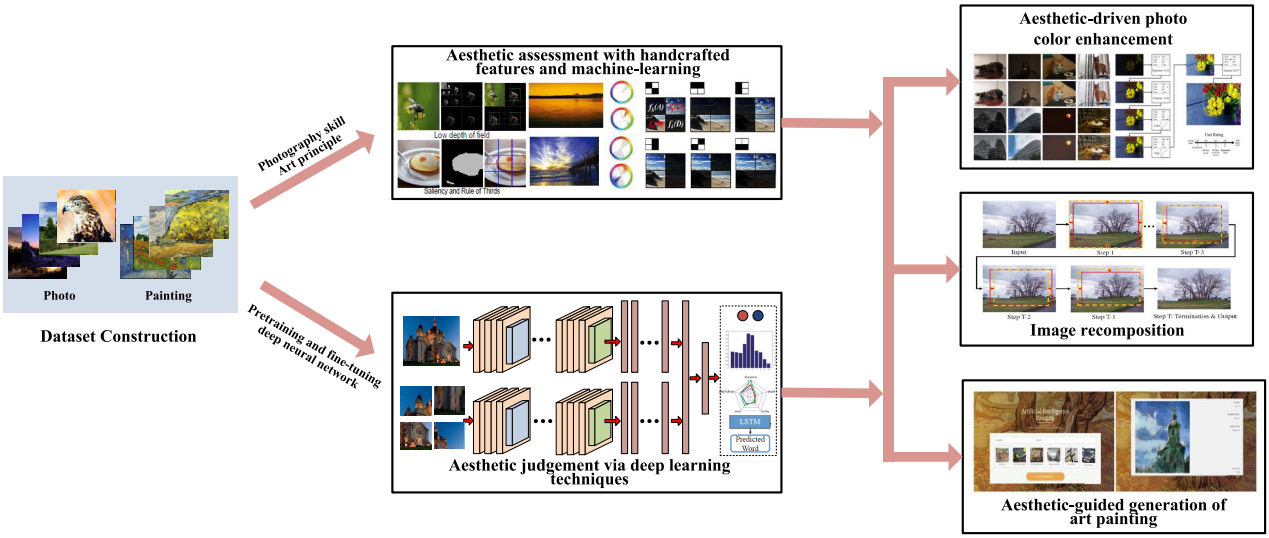


FIGURE 2. The framework for research on computational aesthetic evaluation of visual art images.

The framework of review on computational aesthetic evaluation of visual art images is shown in Figure 2. Specifically, we review the most commonly used publicly available aesthetic assessment datasets on different categories of art images. Then, conventional aesthetic evaluation approaches based on handcrafted features are summarized. Besides, we systematically evaluate recent deep learning techniques that are useful for developing robust models in aesthetic judgment on scoring, distribution, attribute, and description, as shown in Figure 3. Moreover, the possibility of building connections between image aesthetic evaluation and aesthetic-driven photo enhancement, including automatic image color adjustment and recomposition, and automatic generation of aesthetic-guided art paintings through computational approaches are analyzed. We hope that our survey could serve as a comprehensive reference beginning for future research on the computational aesthetics in visual media and its valuable potential applications.

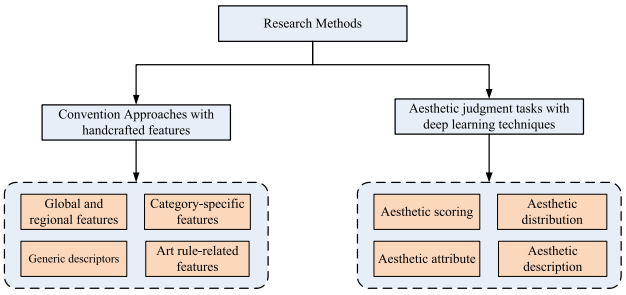


FIGURE 3. Research methods classification.

The remainder of this paper is organized as follows: We first give a review of image aesthetic assessment datasets in Section II. Then in Section III, we summarize conventional approaches based on handcrafted features. The aesthetic evaluation tasks with deep learning are analyzed in Section IV. In Section V, we build a connection between aesthetic assessment and aesthetic manipulation, with a focus on aesthetic-based image enhancement and automatic art painting generation. In Section VI, current challenges of research and some open issues for future directions in this field of study are discussed. Finally, we conclude our survey in Section VII.

II. DATASETS

Being treated as a data-driven learning problem, the construction of the image aesthetic evaluation benchmark dataset has become the key prerequisite for the research. Many attempts have been made to contribute publicly available large-scale datasets for more standardized evaluation of model performance. In the acquisition of subjective scores of image aesthetics, it can be realized through manually scoring experiments in the lab [38]–[40], online scoring on image sharing website [26], [41], and crowdsourcing evaluation [42], [43]. In the following, we introduce some most commonly used benchmark datasets for image aesthetic assessment in photographs and paintings, respectively.

A. NATURAL PHOTOGRAPH IMAGE

The Photo.Net dataset [26] is one of the earliest large-scale databases for image aesthetic evaluation. It contains 20,278 images with at least 10 ratings per image, each of which ranges between 1 and 7 with 7 represented the highest aesthetic feeling, and the overall distribution is skewed towards aesthetically pleasing. Tang *et al.* [38] constructed the CUHK-Photo Quality (CUHK-PQ) dataset, which contains more than 17,690 images with binary aesthetic labels assigned by at least 8 out of 10 subjects for each sample. All images are grouped into 7 semantic categories including "animal", "plant", "static","architecture","landscape",

"human", and "night", with the overall ratio of the total number of positive and negative samples around 1: 3. The Hidden Beauty of Flickr Pictures (HiddenBeauty) dataset [43] contains more than 15,000 images belonging to one of four categories including "human", "nature", "urban", and "people", which were chosen from the YFCC100M dataset [54]. The aesthetic score of each image was collected on a five-point scale through the CrowdFlower crowdsourcing platform.

Murray *et al.* [41] built the first large-scale database for aesthetic visual analysis (AVA) dataset containing 255,530 images with detailed annotation. The images are downloaded from www.dpchallenge.com, an online image sharing and scoring website storing a community of professional and amateur photographers. Each image in the dataset is associated with a distribution of scores representing individual votes, which are rated on a scale 1-10 by receiving between 78 and 549 votes by amateur and professional photographers, and the mean score is calculated to be the ground truth aesthetic label for the image, with the overall ratio of the total number of positive and negative samples around 12:5, as shown in Figure 4. Along with aesthetic scores, the dataset includes 66 semantic attributes and 14 photographic style annotations (e.g. Complementary colors, High Dynamic Range, Motion Blur, etc.). The AVA dataset serves as an acknowledged benchmark for performance evaluation in image aesthetic evaluation and aesthetic distribution learning. Schwarz *et al.* [44] built a large-scale multi-user agreements image aesthetic dataset (AROD), which contains 380K images downloaded from the online image sharing website Flickr with associated metadata such as the number of views, favorite list, etc. To annotate the aesthetic score of an image, they calculate the ratio between the number of views and the number of clicks that favor the image to quantify the human aesthetic feeling.

FIGURE 4. (a) Sample images in the AVA dataset. Images rated with mean score > 5 are grouped in green, while the ones rated with mean score < 5 are grouped in red. The right groups are ambiguous ones with moderate scoring around 5. (b) The distribution of positive and negative images for different partitions in AVA dataset [33].

Besides these standard benchmarks, there are some new datasets that solving the balanced distribution in different aesthetic levels with richer high-level aesthetic attributes and description annotations. Kong *et al.* [45] constructed a new Aesthetics and Attributes database (AADB), which contains totally 10,000 images with overall aesthetic scores and binary labels of 8 aesthetic attributes (balancing element, color harmony, interesting content, shallow depth of field, good lighting, object emphasis, rule of thirds, vivid color) rated by 5 different workers through Amazon Mechanical Turk (AMT), as shown in Figure 5. Chang *et al.* [46] issued the first aesthetic captioning dataset called Photo Critique Captioning dataset (PCCD), which contains 4,235 images and 29,645 pairwise aesthetic comments from professional photographers in 7 aspects, including general impression, subject of photo, composition, and perspective, depth of field, color and lighting, focus, and use of camera, exposure and speed, as shown in Figure 6. The AVA-Comments [48] and AVA-Reviewers [47] datasets are designed by selecting 255,530 and 40,000 images from AVA dataset and add single sentence comments to describe overall impression. Jin *et al.* [49] built a new dataset named DPC-Captions. It contains 154,384 images and 2,427,483 comments of up to 5 aesthetic attributes using aesthetic knowledge transfer from the full-annotated small-scale PCCD and AVA-Plus datasets, which crawled 330,000 images together with their comments from DPChallenge.com to a large-scale weakly annotated one.

FIGURE 5. Sample images in the AADB dataset. Each photo is annotated with the 8 aesthetic attributes in binary labels and aesthetic ratings on a scale 1 to 5 (displayed on top and right of each image, respectively) [45].

Table 1 lists a summary comparison between different public datasets for photo aesthetics assessment. The AADB and AVA provide a larger scale, broader score distribution, richer semantic and style attribute annotations than Photo.Net and CUHK-PQ, which are either biased or consist of samples for easy binary aesthetics classification. The key differences between AADB and AVA are that in AVA many images are heavily edited or synthetic, while AADB contains a much more balanced distribution of photographic imagery of real scenes downloaded from Flickr. However, the quantity is small and the tag of each aesthetic attribute in AADB is a binary value (high or low aesthetics). The PCCD and DPC-Captions provide more detailed comments for each aesthetic attribute than AADB and AVA. While the size of PCCD is relatively small compared to AVA, which is commonly used in this field but does not contain ground truth of aesthetic captions and attributes. The AVA-Reviews

TABLE 1. Comparison of the properties in different benchmark datasets on image aesthetic assessment.

Dataset	Number of Images	Scale	Category	Score distr.	Average raters	Attributes	Comments	Annotations
Photo.Net [26]	20,278	1-7	–	Positive bias	10	–	–	Aesthetic score
CUHK-PQ [38]	17,690	0-1	7	Negative bias	8	–	–	Binary label
HiddenBeauty [43]	15,000	1-5	4	–	5	–	–	Aesthetic score
AVA [41]	255,530	1-10	66	Positive bias	210	14	–	Distribution
AROD [44]	380,000	–	–	Uniform	6868	–	–	Aesthetic score
AADB [45]	10,000	1-5	–	Normal	5	8	–	Binary Attribute
PCCD [46]	4,235	1-10	–	–	7	7	29,645	Comments
AVA-Reviews [47]	40,000	1-10	66	Positive bias	6	–	240,000	Comments
AVA-Comments [48]	255,530	1-10	66	Positive bias	6	–	1,535,937	Comments
DPC-Captions [49]	154,384	1-10	–	–	15	5	2,427,483	Comments
Oil painting [39]	100	1-5	1	Positive bias	42	4	–	Aesthetic score
JenAesthetics [50]	1628	1-100	16	Positive bias	20	7	–	Aesthetic score
Chinese painting [51]	511	1-9	2	–	20	5	–	Aesthetic score
Abstract painting [40]	1000	1-7	–	Positive bias	20	–	–	Emotion score
WikiArt Emotions [52]	1000	1-20	22	Positive bias	10	6	–	Emotion label
ArtEmis [53]	81,446	1-8	45	Positive bias	5	–	439,121	Comments

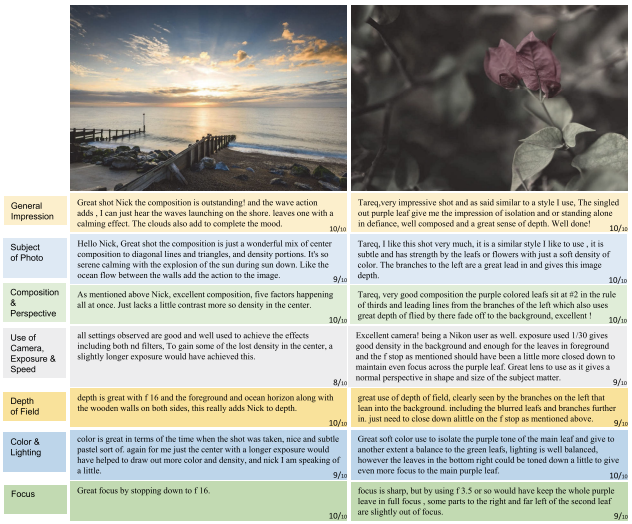


FIGURE 6. Samples in the photo critique captioning dataset with scores and comments in 7 aesthetic aspects [46].

and AVA-Comments do not annotate the comments for individual aesthetic attributes. Although DPC-Captions has fewer images than AVA-Comments, the average number of comments for each image in DPC-Captions is larger than that of AVA-Comments. The small-scale PCCD contains both comments and scores of attributes, while the large-scale DPC-Captions only contains partially annotated attribute comments.

B. PAINTING ART IMAGE

The aesthetic annotation of painting art images is usually obtained through aesthetic experiments, such as gathering non-art subjects in a controlled environment and recording the participants' aesthetic responses to the paintings combined with discrete vocabulary and hierarchical aesthetic representation method.

Li and Chen [39] collected 100 oil painting images of impressionistic landscape as subject matter, and invited 42 non-art volunteers to score each oil painting (1-5 points) in 4 dimensions including color, composition, texture, and general aesthetic feeling, and the average scores are used in aesthetic quality evaluation of oil painting, as shown in Figure 7. Amirshahi *et al.* [50] built a publicly available JenAesthetics dataset, which contains 1,628 high-quality images of colored oil paintings downloaded from the Wikimedia Commons website in Western provenance from 410 artists. The art images in the dataset cover a wide range of 11 art periods (e.g. Classicism, Realism, Rococo, etc.) and 16 subject matters (e.g. Abstract, Urban scene, Still life, etc.). For each session in the human rating, 163 images are selected randomly from the dataset and presented to the observer, and each painting is rated on a scale 1-100 between 19 and 21 votes from non-art observers. The subjects are asked to score each image on 7 properties including "Aesthetic quality", "Beauty", "Liking of color", "Liking of content", "Liking of composition", "Knowing the artist", and "Familiarity with the painting", using a sliding bar located on the bottom of the screen, and the median value between the scores in each aspect is calculated as the final ground truth label. Zhan *et al.* [55] designed a new dataset for aesthetic and emotional evaluations of traditional Chinese paintings. The dataset contains 511 images of Chinese paintings collected from multiple sources

FIGURE 7. Sample images in the dataset of impressionistic landscape oil paintings that are labeled as "high-quality" (the upper row) and "low-quality" (the bottom row) [39].

(e.g. www.artsjk.com, an online Chinese art website). 350 aesthetic adjectives are then collected and filtered as 5 aesthetic semantic categories through Hevner affective ring, questionnaire survey, and factor analysis. 20 participants are asked to make ratings for each image on 5 items: aesthetic category including beauty score of momentum, quiet, vitality, elegant, bleak, and pleasure-arousal-dominance (PAD) on a scale 1-9.

Sartori *et al.* [40] built a new dataset for emotion recognition evoked by collecting 500 professional abstract artworks created by 78 artists from the MART dataset, and 500 amateurs abstract artworks created by 406 authors from the deviantArt dataset, an online social network website sharing user-generated artworks, as shown in Figure 8. To collect the ground truth of positive and negative emotions evoked by abstract paintings, they used an Absolute Scale annotation, in which 100 raters including visitors, teachers, curators, and students were asked to use the Likert scale of 1-7 points to judge paintings presented one by one, where 1 meant a highly negative emotion and 7 meant a highly positive emotion. Besides, the relative scale annotation was used to collect the ground truth for comparison. 25 subjects were presented with paintings in pairs and were asked to select the one which evoked more positive emotions. Then the TrueSkill ranking system was used to get a reliable ranking by optimally sampling pairs of paintings, the final ''skills'' were considered as emotional scores. Mohammad and Kiritchenko [52] created the WikiArt Emotions Dataset, which contains emotion annotations for 4,105 pieces of art selected from WikiArt.org's collection for 22 categories (e.g. impressionism, realism, etc.) in 4 western styles (Modern Art, Post-Renaissance Art, Renaissance Art, and Contemporary Art). Each art image is annotated by crowdsourcing for 20 emotions classified in ''positive'', ''negative'', and ''other or mixed'', whether it shows the depiction of a face, and how much the observers like the art (-3-3 points). The distribution showed that about 64% of the images were labeled as liked, 18% as disliked, and 18% as neither liked nor disliked. Achlioptas *et al.* [53] present a new large-scale dataset providing affective explanations for the interplay between visual content and its emotional effect, named ArtEmis. It contains 81K artworks from WikiArt covering 27 art styles (abstract, cubism, impressionism, etc.) and 45 genres (cityscape, landscape, portrait, etc.). Each artwork was annotated by at least 5 subjects to express their dominant emotional reaction by choosing among the 8 emotions (Amusement, Anger, Disgust, Excitement, etc.), and explaining the reason for their response, which are 439,121 emotion explanations in natural language from visual stimuli in the artworks, as shown in Figure 9.

The comparative statistics of different public datasets for aesthetics assessment of painting arts are shown in Table 1. The ArtEmis provides a larger scale, greater diversity in category distribution of art styles and genres, and richer semantic explanations to elicit emotional responses in artworks than abstract painting and WikiArt Emotions. The Jenaesthetics is more focused on annotation in aesthetic feeling and

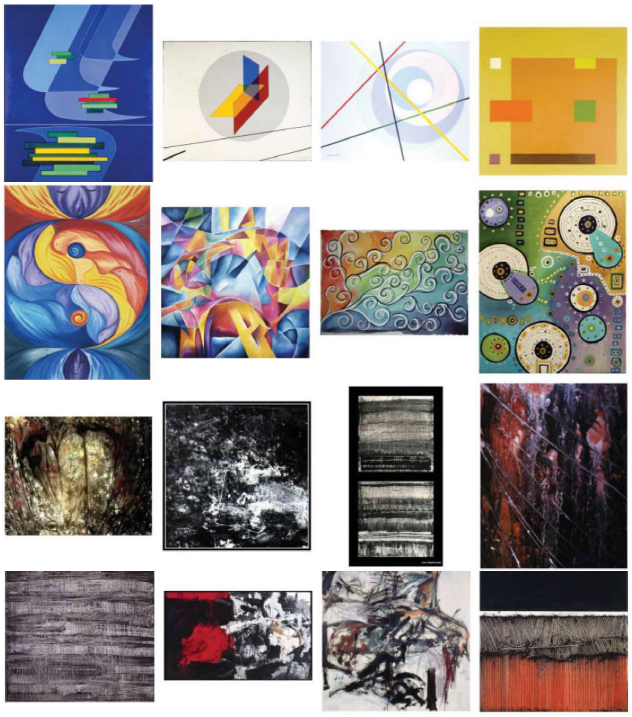


FIGURE 8. Sample images in the dataset of abstract paintings from MART and deviantArt that are classfied as "highly positive" (the upper two rows) and "highly negative" (the bottom two rows) [40].

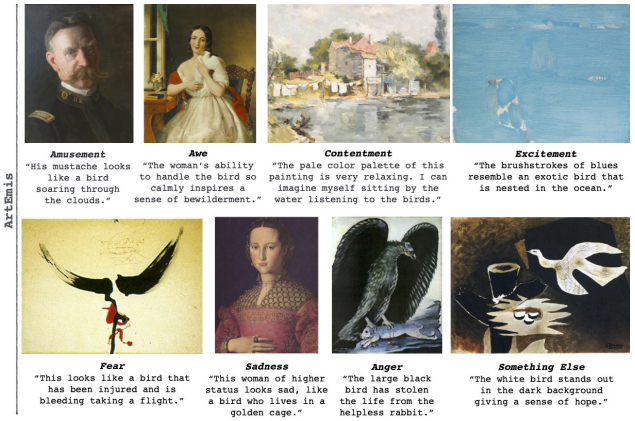


FIGURE 9. The affective explanations in ArtEmis (top and middle rows) explain abstract semantics and emotional states that are not directly visible associated with the contents [53].

associated aesthetic factors in western paintings, which is more comprehensive than the oil painting dataset.

III. CONVENTIONAL EVALUATION WITH HANDCRAFTED AESTHETIC FEATURES

The conventional approach for aesthetic evaluation of visual art images is to extract hand-designed features, which quantifies various critical photography skills and domain expertise in painting art rules. Below we summarize a variety of approaches in aesthetic evaluation of photo images in content-independent features, category-specific features, and

TABLE 2. Comparison of various conventional evaluation approaches with handcrafted aesthetic features.

Type	Literature	Main idea and contributions
Global and regional features	Ke *et al.* [15](2006)	– Modeling the global aesthetic properties of photos
	Sun *et al.* [56](2009)	– Using global saliency map to estimate visual attention distribution
	Aydin *et al.* [57](2015)	– Computing meaning aesthetic signatures to predict the aesthetic score
	Luo *et al.* [16](2008)	– Designing semantic features based on subject and background regions division
Category-specific features	Dhar *et al.* [17](2011)	– Using low-level features to estimate high-level human-describable attributes
	Tang *et al.* [38](2013)	– Develop new subject area extraction methods and visual features for different categories
	Rawat *et al.* [58](2015)	– Develope a scene-dependent photography model associated with contextual features
	Lien *et al.* [59](2015)	– Compute face-specific aesthetic features
Generic descriptors	Marche *et al.* [60](2011)	– Using generic signature to encode aesthetic properties
	Nishi *et al.* [61](2011)	– Develop a bag-of-features framework based on color harmony
	Su *et al.* [62](2012)	– Construct an aesthetic feature library in preference-aware aesthetic model
Art rule-related features	Li *et al.* [39](2009)	– Extracting characteristics related to artistic knowledge in oil paintings
	Mallon *et al.* [63](2014)	– Establish the correlation of beauty rating and perceptual contrast with PHOG
	Sartori *et al.* [64](2016)	– The influence of color and brush stroke on emotion evoked by abstract painting
	Zhang *et al.* [65](2017)	– Develop an aesthetic model of Chinese ink paintings with 7 low-level art features

painting images based on art rules and cognitive psychology (Figure 10), as shown in Table 2.

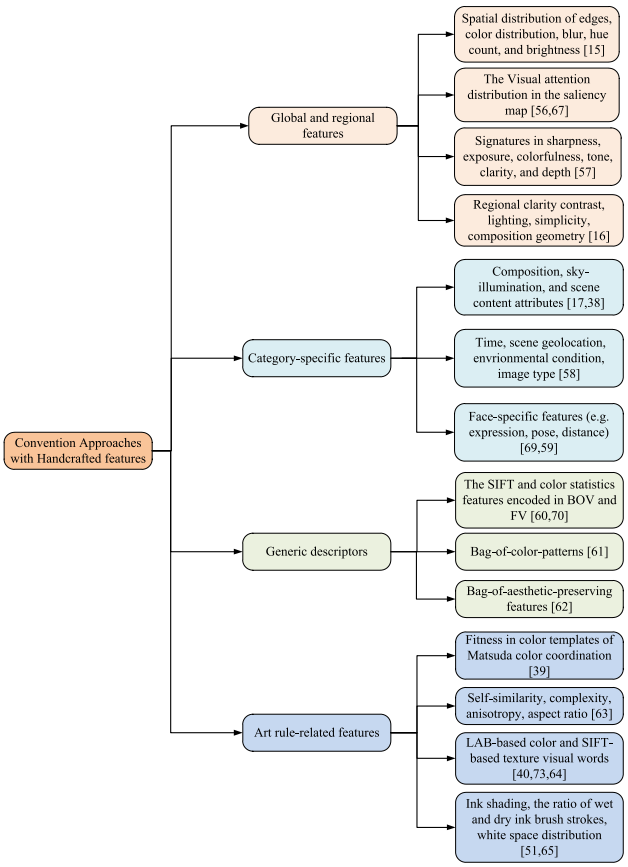


FIGURE 10. Convention approaches with handcrafted features.

A. GLOBAL AND REGIONAL FEATURES

By using simplicity, realism, and other basic photography techniques as guidelines, early works extracted global features to implicitly model specific photographic rules or aesthetic principles. Ke *et al.* [15] is one of the first attempts in quantitatively modeling the global aesthetic properties of photo images using hand-designed features. They extract spatial distribution of edges, color distribution, blur, hue count, and brightness to differentiate between professional photographs versus snapshots. Sun *et al.* [56] trained a regression model based on a global saliency map to estimate visual attention distribution, they calculate the rate of focused attention region in the saliency map to predict the aesthetic quality score of an image. Encapsulated aesthetic signatures including sharpness, exposure, colorfulness, tone, clarity, and depth were computed from images in [57], which comprise calibrated ratings of meaningful attributes to predict the overall aesthetic evaluation score.

Later some works extracted the foreground region focused on the subject, and the regional features were then calculated to be effective in complementing the global features for improving the model classification performance. Luo and Tang *et al.* [16] used the blurred region detection algorithm to extract the subject region and the rest as the background from the photo image, and then extracted regional clarity contrast, lighting, simplicity, composition geometry, and color harmony features based on this subject and background division. Wong and Low [66] proposed a saliency map model to classify the image aesthetic quality. They first adopt a visual attention model to extract the saliency regions in the photo, then the exposure, sharpness, and texture features on global image and salient regions, as well as features describing the relationship of subject and background are computed from the image.

B. CATEGORY-SPECIFIC FEATURES

While all the above approaches attempt to train the universal aesthetic models to deal with all types of photos without considering the differences in photo content. Since professional photographers utilize different photography skills and have different aesthetic criteria when capturing different types of photos, some works exploited category-specific features (such as object and scene types, sky illumination, portrait face attributes, geographic location characteristics, etc.) based on the different task nature in image aesthetic assessment.

Dhar *et al.* [17] used low-level features to estimate high-level human-describable attributes, such as composition, sky-illumination, and scene contents, which were used to predict aesthetic quality and interesting of photos. Specifically, for the content attributes, they aim at the presence of people and portrait depiction, presence of animals, indoor-outdoor, 15 other different scene categories and trained corresponding classifiers. For the sky-illumination attributes, they focus on training classifiers of natural outdoor illumination in clear skies, cloudy skies, and sunset skies. Tang *et al.* [38] divided the photos into 7 categories including ''animal'', ''plant'', ''static'', ''architecture'', ''landscape'', ''human'', and ''night'' based on different visual contents, and a set of new subject area extraction methods and visual features were specially designed for different categories. Rawat and Kankanhalli [58] developed a scene-dependent photography model that can help an amateur in capturing high aesthetics photographs. They incorporate the learning with associated contextual features such as time, geolocation that links a target image to relevant photos within the same scene geo-context, environmental conditions, and image types, and composition information such as eigenrules and baserules. The proposed model can be used to provide guidance to the user concerning scene composition and camera parameters while capturing photos.

Li *et al.* [67] proposed a photo quality assessment and selection system based on aesthetic evaluation of consumer photos with faces. Face-specific aesthetic features (such as individual facial expressions, individual face poses, and between-face distances) are specifically captured and complemented with the conventional features (color, lighting contrasts between the face and background regions, composition rules in faces for assessing and improving the quality of portraiture. Lienhard *et al.* [59] divided the input headshot image into different regions (global face region, eyes region, and mouth region) and computed 15 low-level features such as illumination, contrast, or colorfulness for each region, then relevant features and facial areas are chosen by a feature ranking algorithm, increasing both model classification and regression performance.

C. GENERIC DESCRIPTORS

An alternative approach, proposed by Marchesotti *et al.* [60], [68] is to explore generic image signatures such as Bag-of-Visual-words (BOV) [69] and Fisher Vector (FV) [70], which have been successfully used for object recognition, to train aesthetics models. The SIFT and color statistics features are used as the local descriptors upon the second order of the Gaussian Mixture distribution and spatial pyramid, which are then encoded using BOV or FV, and concatenated as the final image representation. The authors prove that generic descriptors can implicitly encode the photographic aesthetic properties with better performance than the existing hand-designed aesthetic features.

Nishiyama *et al.* [61] decomposed a photograph into a collection of local regions with color variations using a grid-sampling technique, as shown in Figure 11(a), and a color harmony model was applied to each local region of an image to compute local descriptors with/without color boundaries to generate codebooks, which were used to evaluate the distributions of dominant color in the region, and then the histograms are integrated to represent the whole image in the bag-of-color-patterns framework, as shown in Figure 11(b). A preference-aware view recommendation system for scenic photos was proposed by Su *et al.* [62] to suggest views according to varied user-favorite photographic styles, where a bottom-up method is designed to construct an aesthetic feature library with bag-of-aesthetics-preserving features instead of top-down approaches that implement the heuristic rules. However, generic features may be unable to attain the upper performance limits in aesthetics-related problems [22].

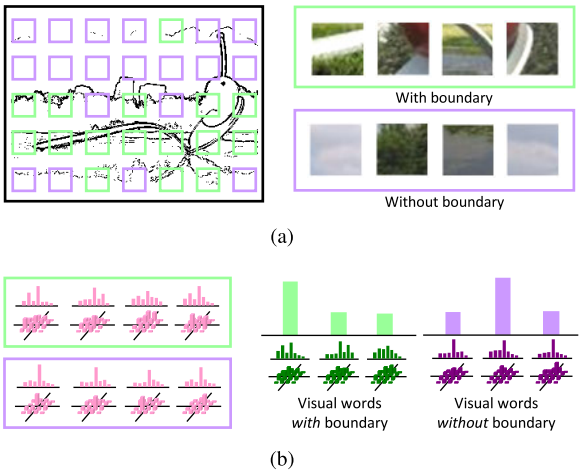


FIGURE 11. The bag-of-color-patterns framework. (a) The sampled local regions with/without color boundaries in two sets; (b) The local descriptors are computed from local regions in (a) to generate visual words in the codebooks, which are used to plot and concatenate the two histograms to deduce a feature representation for the photo [61].

D. ART RULE-RELATED FEATURES

In terms of western paintings, Li and Chen [39] detected some characteristics related to artistic knowledge, such as color distribution that fitting into the 8 hue and 10 tone types of Matsuda's color coordination, brightness, blur effect, and edge distribution globally, together with local features such as the shape of regions and color contrast features between different segments, to classify the impressionist oil paintings by Van Gogh and Monet as high or low aesthetic quality. Mallon *et al.* [63] established the correlations of beauty ratings and perceptual contrast with statistical properties of abstract artworks. They first investigate shifts in aesthetic perception when appreciating abstract artworks, which revealed a clear pattern of perceptual contrast. Then they extract the Pyramid of Histograms of Orientation Gradients(PHOG) such as self-similarity, complexity, anisotropy, and aspect ratio as predictive variables to construct the aesthetic classification model. Sartori *et al.* conducted a series

of emotional researches on abstract paintings, including distinguishing the difference between positive and negative emotional abstract paintings by using the general bag-of-visual-words classification framework, which extracted LAB-based color visual words and SIFT-based texture visual words [40], the influence of color combination on emotion [71], visual and metadata features for improving the abstract painting emotion identification performance [72], and how color and brush stroke in abstract paintings could be used to build a computational model that predicts whether an abstract painting would elicit positive or negative emotions in the observer [64]. Amirshahi and Denzler [73] designed features include color self-similarity, weighted color self-similarity, appearance color heterogeneity, and other characteristics for aesthetic classification of western paintings.

For aesthetic assessment of Chinese paintings, Zhang *et al.* [65] developed an aesthetic model of Chinese ink paintings with 7 low-level handcrafted features such as ink shading, the ratio of wet and dry ink brush strokes, selected through stepwise linear regression. Unfortunately, the variation effect of line rules and white space distribution are ignored in this model. Besides, the dataset is built by collecting human ratings of overall feeling, color collocation, composition, and brush stroke in only 60 flower and bird paintings from Qi Baishi, which results in lacking generalization in the aesthetic model. Zhan *et al.* [51] conducted the automatic aesthetic classification of Chinese paintings based on the 5 types of aesthetic annotation in the database. 33 image features suitable for aesthetic classification are obtained, and the results show that the art elements related to the aesthetic feeling of Chinese paintings are ranked in order of importance: color, brush strokes, brightness, and lines.

While these handcrafted aesthetics features achieved good evaluation performances, they have some limitations: First, the manually designed aesthetic features based on specific photographic criteria have a limited range, it is impossible to cover exhaustive effective photographic attributes. Second, due to the vagueness of certain photographic or psychologic rules and the difficulty in evaluating them quantitatively, those handcrafted features are usually just approximations of specific rules. We would summarize some recent literature in aesthetic assessment using the deep learning technique in the following sections.

IV. AESTHETIC JUDGEMENT WITH DEEP LEARNING APPROACHES

Beginning with the strong performance of Krizhevsky *et al.* [18] in the image classification, the powerful feature representation learned with a growing amount of datasets, and feasible transfer learning [19] with fine-tuned Convolutional Neural Networks (CNN) [20], deep learning methods [21] have been applied to aesthetic quality assessment of visual art images, which can automatically learn effective aesthetic features from deep hidden layers to abstract image information without expert knowledge, thus showing outperformed evaluation capability than conventional handcrafted features. Researches in the aesthetic assessment of visual art images using deep learning approaches can be summarized into 4 major schemes (Figure 12), 1) Aesthetic scoring refers to different aesthetic quality levels (binary labels as ''positive'' or ''negative''), or continuous images aesthetic ratings; 2) Aesthetic distribution refers to the distribution histogram of aesthetic scores of images; 3) Aesthetic attribute refers to the evaluation of good lighting, color harmony, shallow depth of field, balancing element, motion blur and other aspects of images; 4) Aesthetic description refers to linguistic aesthetics comments of images, as shown in Table 3.

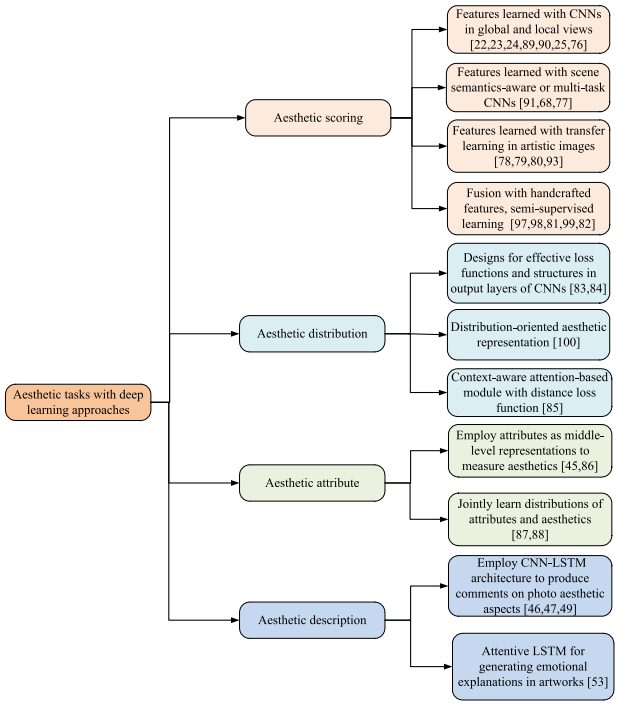


FIGURE 12. Aesthetic tasks with deep learning approaches.

A. AESTHETIC SCORING

1) FEATURES LEARNED WITH CNNs IN GLOBAL AND LOCAL VIEWS

The RAPID model proposed by Lu *et al.* [22] can be considered as the first attempt in training convolutional neural networks for aesthetic quality classification. They use an AlexNet-like architecture where the last fully-connected layer is set to output 2-dim probability for aesthetic binary classification. The best model is obtained by stacking a global warped image and a local random cropped patch as inputs to form a double-column CNN, where the feature representation from each column is concatenated before the classification layer, as shown in Figure 13. Moreover, they further improve the performance of the network by incorporating image style information using a style-column CNN as the third input column, forming a three-column CNN with semantic information. However, a single patch may not always well represent

TABLE 3. Comparison of different aesthetic judgment tasks with deep learning approaches.

Types	Literatures	Main ideas and contributions
Aesthetic scoring	RAPID [22](2014)	– Develop a double-column architecture to jointly learn features from heterogeneous inputs – Improve aesthetic categorization using style attributes
	DMA-Net [23](2015)	– Propose a deep multi-patch aggregation network training approach
	A-lamp [24](2017)	– Present an Adaptive Layout-Aware Multi-Patch CNN architecture
	MNA-CNN [74](2016)	– Propose a composition-preserving deep Multi-Net Adaptive Spatial pooling CNN
	GPF-CNN [25](2019)	– Design a double-subnet Gated Peripheral-Foveal CNN with a gated information fusion module
	MSDLM [75](2016)	– Design a multi-scene convolutional layer to make the network a strong adaptability
	MTCNN [76](2017)	– Exploit semantic recognition information in aesthetic representation with a multi-task CNN
	Li *et al.* [77](2017)	– Transfer learning of bilinear CNN model pretrained in natural images to sketch works
	Zhang and Xu [78](2019)	– The pre-trained VGG16 model is fine-tuned on the ethnic paintings for emotional classification
	Inkthetics [79](2020)	– Design a comprehensive deep multi-view parallel CNN for aesthetics in Chinese paintings
	MRACNN [80](2020)	– Propose a novel multimodal recurrent attention CNN with vision and language streams
	Sheng *et al.* [81](2020)	– Propose a self-supervised learning scheme without manual annotations
Aesthetic distribution	NIMA [82](2018)	– Using Earth Mover's Distance to measure the distance of real and predicted histograms
	CJS-CNN *et al.* [83](2018)	– Measure the cumulative distribution loss function with Jensen-Shannon divergence
	Xu *et al.* [84](2020)	– Explore a context-aware attention-based model to predict human opinions distribution
Aesthetic attribute	Kong *et al.* [45](2016)	– Adding an attribute prediction branch into the basenet to fuse with the aesthetic branch
	DCM [85](2017)	– Learn attributes through the parallel supervised pathways into the overall aesthetics
	Viswa *et al.* [86](2020)	– Jointly learn aesthetic attributes along with the overall aesthetic score in a multi-task CNN
	Pan *et al.* [87](2019)	– Using adversarial learning to model the joint distributions of aesthetics and attributes
Aesthetic description	Chang *et al.* [46](2017)	– From judgement to critiques with multi-aspect aesthetic captioning
	Jin *et al.* [49](2019)	– Design Aesthetic Multi-Attribute Network to produce both captions and scores for attributes
	Achlioptas *et al.* [53](2021)	– Train affective neural speakers that produce emotional explanations in artworks

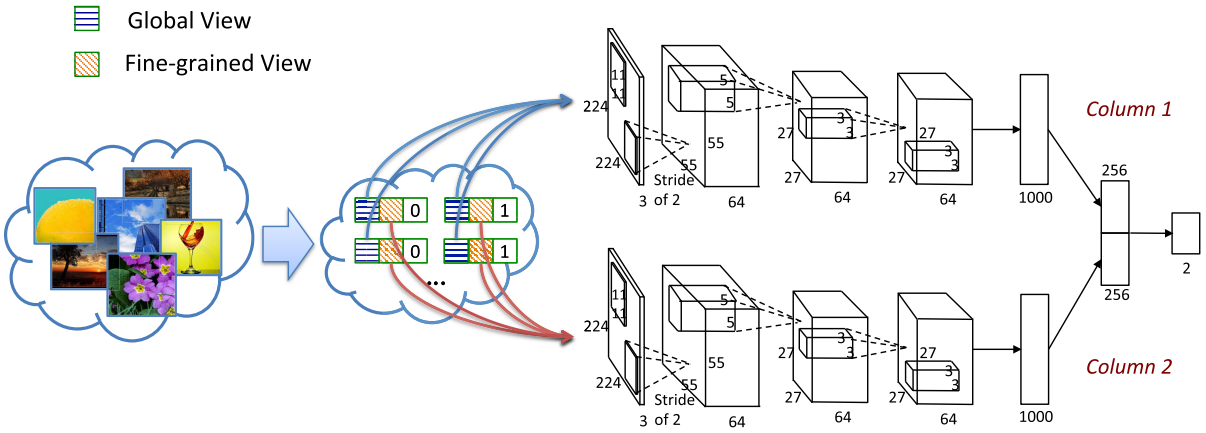


FIGURE 13. Double-column convolutional neural network for rating pictorial aesthetics. Each training sample is represented by its global and local views, associated with its aesthetic quality label: low quality (0) and high quality (1). The final fully-connected layer are jointly trained [22].

the aesthetic properties of the entire image. Then they propose a deep multi-patch aggregation network architecture (DMA-Net) [23], which could train models using multiple patches generated from one image. It is achieved by constructing multiple shared CNNs and feeding multiple patches to each of the columns. More importantly, a statistical aggregation structure is designed to aggregate the features from the orderless sampled patches by different poolings (min, max, median and averaging), and an alternative aggregation structure is also designed based on sorting.

While the DMA-Net has shown some promising evaluation capability, it takes multiple randomly cropped patches as inputs, which are unable to capture the image layout information. Ma *et al.* [24] developed an Adaptive Layout-Aware Multi-Patch CNN (A-Lamp CNN) architecture, which extracts features from both fined-grained details and holistic image layout simultaneously. To support training on these hybrid inputs, they design a double-subnet neural network structure including a Multi-Patch subnet and a Layout-Aware subnet. For an arbitrary-sized input image, multiple patches are adaptively selected based on several criteria including saliency map, pattern diversity, and overlapping constraint instead of random-cropping, and fed into the Multi-Patch subnet. A statistic aggregation structure is followed to effectively combine the feature representations from these CNNs, as shown in Figure 14. Meanwhile, the global layout of the input image is further expressed via attribute graphs. Finally, a learning-based layer is used to aggregate the hybrid features from the two subnets in predicting aesthetics. However, the number of attribute graph node need to be predefined. Chen *et al.* [88] proposed a double-column CNNs for learning aesthetic feature representation, which

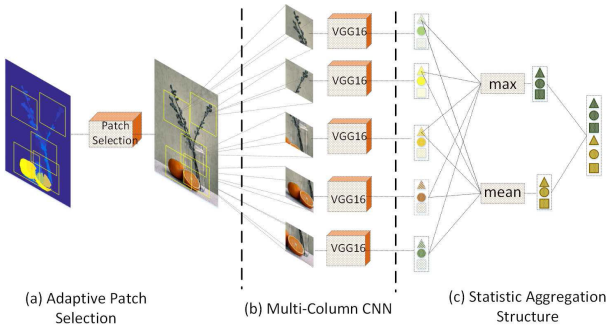


FIGURE 14. The architecture of Multi-Patch subnet in A-lamp CNN: (a) adaptive patch selection module. (b) a set of paralleled CNNs that are used for extracting deep features. (c) aggregation structure that combines the extracted deep features from the multi-column CNNs jointly [24].

uses a weakly-supervised learning algorithm to project a set of textual attributes learned from image labels to highly responsive image regions. Such patches in images are then fed to the Multi-Patch CNN, with a parallel output branches modeling each one of the textual attributes are concatenated on top. Sheng *et al.* [89] presented a novel multi-patch aggregation method for image aesthetic assessment. They design 3 attention-based objective functions (i.e., average, minimum, and adaptive) that adaptively adjusted the weight of each patch to enhance the training efficiency.

Inspired from the mechanism in human aesthetic perception, Zhang *et al.* [25] presented a double-subnet Gated Peripheral-Foveal Convolutional Neural Network (GPF-CNN), which simulates the peripheral vision to encode the holistic information and provided the attention regions, and the foveal vision to extract fine-grained features on the attended regions. Considering that peripheral vision and foveal vision play different roles in processing different visual stimuli, they further develop a gated information fusion module to adaptively balance the weights of the global and local subnets, which are determined through the fully connected layers followed by a sigmoid function. The comprehensive experiments for different tasks in aesthetic classification, aesthetic regression, and aesthetic distribution show that the proposed model outperforms the state-of-the-art methods on the AVA and Photo.net datasets.

From another perspective, considering the damage of image aesthetics due to the fixed-size input restriction, Mai *et al.* [74] proposed a composition-preserving deep Multi-Net Adaptive Spatial Pooling Convolutional Neural Network (MNA-CNN) architecture for photo aesthetic assessment that can directly process the original input images without any image transformation. Specifically, they add an adaptive spatial pooling layer upon the regular convolution and pooling layers to directly handle input images with arbitrary sizes and aspect ratios. To encode multi-scale image information, their architecture consists of multiple subnetworks, each having an adaptive spatial pooling layer with a different pooling scale. Moreover, a scene-aware aggregation layer is designed as category posterior to effectively combine the multi-scale VGG features from sub-networks. The advantage of the MNA-CNN is to capture aesthetics features at multiple scales, yet the multi-scale VGG features may contain redundant information and lead to model overfitting [33].

2) FEATURES LEARNED WITH SCENE SEMANTICS-AWARE OR MULTI-TASK CNNs

Different scenes tend to employ different photographic techniques and aesthetic criteria. Therefore, a universal aesthetic model cannot always capture the full diversity of scene semantics. By integrating the scene semantic information into the deep neural network, the aesthetic predictive performance has been significantly improved. Tian *et al.* [90] proposed a query-dependent photo aesthetic evaluation model based on the efficient feature representation learned from a single-column CNN. Specifically, for a given query image, a query-dependent training set is retrieved based on visual similarity, image text tags, or the fusion of both. Then the SVM classifier is trained on this retrieved training set. The experiments show that the query-dependent model outperformed the universal model learned in the ImageNet task across all 7 categories.

Wang *et al.* [75] presented a multi-scene deep learning CNN model (MSDLM) that is modified from the AlexNet architecture to comprehensively learn aesthetic features. To improve the network adaptability to different scenes, they replace the $conv_5$ layer of AlexNet by designing a scene convolutional layer, which contains 7 convolutional groups paralleled linked to $conv_4$ by 7 independent branches for different scene categories ($conv_5^1 - animal$, $conv_5^2 - architecture$, $conv_5^3 - human$, $conv_5^4 - lanscape$, $conv_5^5 - night$, $conv_5^6 - plant$, $conv_5^7 - static$), with mean pooling before feeding into the fully-connected layer, thus could exploiting the aesthetic descriptors discriminatingly according to each of the 7 scene categories. In the pre-training stage, each group is independently trained by using images in one scene category, then the weights learned are paralleled linked back to the previous layer. Then the weight of CNN is further learned through supervised fine-tuning end-to-end. The experiments show that the $conv_5^i$ layer feature map contains a stronger response for the input image of *ith* category, which verifies the comprehensive ability in extracting aesthetic features for different scenes.

Kao *et al.* [76] proposed to exploit semantic recognition information to assist in learning aesthetic representation with a multi-task CNN (MTCNN), as shown in Figure 15. Specifically, they set the aesthetic quality evaluation as the main task and semantic recognition as the aided task. This problem is interpreted as a multi-task probabilistic framework with a bayesian analysis to assess a 2-class aesthetic label and semantic label as a 29-dim binary vector of semantic tags selected in AVA dataset. Then they model the correlation between aesthetic and semantic tasks by designing a multi-task relationship learning framework in controlling the balance of parameters' complexities between the two tasks, which effectively optimize the model parameters and exploit

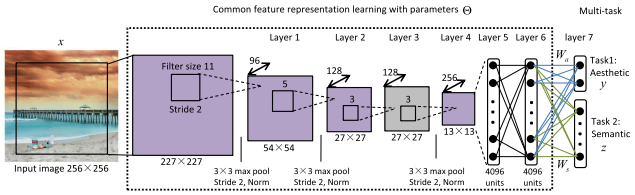


FIGURE 15. The architecture of multi-task convolutional neural network in exploiting semantic recognition information to assist in learning aesthetic representation [76].

the inter-task relatedness for aesthetic feature learning. The experiments show that the multi-task deep framework could explore an effective aesthetic representation with the influence of semantic information with improved accuracy of aesthetic classification.

3) FEATURES LEARNED WITH TRANSFER LEARNING IN ARTISTIC IMAGES

The evaluation based on transfer learning mainly refers to the transfer of natural image related algorithms and knowledge into the aesthetic assessment of artistic images. Sabatelli *et al.* [91] used the VisualBackProp visualization method [92] to observe the deep neural network before and after fine-tuning weights, which shows that the network activation area before fine-tuning is mainly concentrated in the location indicated the object type, and the active area is then moved into the position that is more related to the specific task in art paintings after fine-tuning. Li *et al.* [77] applied deep convolutional features in the classification and evaluation of sketch copy works, which were collected in the teaching scenario including 25 types of works (such as apple, ceramic pot, glass, cube, molectron, etc.), with each category 50 works. They utilize the output of the 6th fully-connected layer in the AlexNet model pre-trained on the ImageNet dataset as deep convolutional feature representation, which is used to train the SVM model in testing the classification performance of sketch works. Then the evaluation task of sketch works is converted as the fine-grained classification problem in 4 levels (best, good, moderate, and worst) based on images' high-level semantic factors (e.g. composition, shape, texture, proportion of black, white or gray). The bilinear CNN model, proposed for fine-grained categorization classification in natural images [93], is fine-tuned end-to-end with the compaction of Tensor Sketch Projection. The results verify that the deep convolutional feature outperformed the traditional features in the classification and aesthetic evaluation of sketch works.

Zhang and Xu [78] researched the classification of positive and negative emotions in ethnic paintings based on the pre-training strategy for related tasks in natural images. The pre-trained VGG16 model on the ILSVRC2012 dataset is fine-tuned on the Twitter Image Dataset to learn effective features for emotional classification of natural images, then the model is trained on the ethnic painting dataset for better adapting on more challenging tasks. To show the learning process of the model more intuitively, they modify the model structure by replacing the last 3 fully connected layers with convolutional layers in different sizes of channels and kernels, with an output of an 8 × 8 sized prediction block to reflect the predicted emotion distribution of the network. Zhang *et al.* [79] proposed a comprehensive framework, named Inkthetics, to quantify aesthetics of Chinese ink paintings based on deep learning, as shown in Figure 16. By establishing an aesthetic assessment dataset, they design a deep multi-view parallel convolutional neural network, which is fine-tuned from the pre-trained VGG16 model by extracting global attribute images and multi-patches as inputs to jointly learn aesthetic features. Moreover, a comprehensive aesthetic evaluation model is trained by fusing the deeply-learned features with handcrafted features that rely on expert knowledge in Chinese paintings.

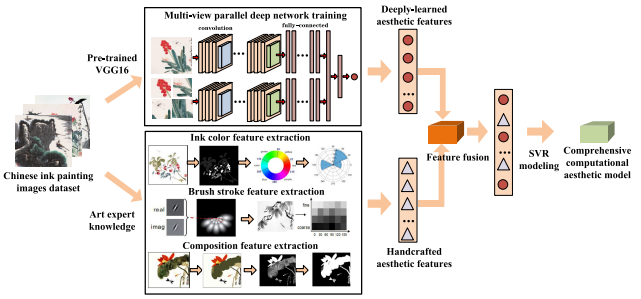


FIGURE 16. The framework of computational aesthetic evaluation of Chinese ink paintings by fusing the deeply-learned features with handcrafted features in expert knowledge in Chinese paintings [79].

Cetinic *et al.* [94] employed deep learning approaches to predict scores related to three subjective aspects of human perception: aesthetic evaluation of the art, sentiment evoked by the art, and memorability of the art. For each concept, they utilize knowledge transfer by evaluating different CNN models trained on various natural image datasets to fine art images. For example, the sentiment recognition model based on Twitter DeepSent [95] and Flickr Sentiment datasets [96], and the aesthetic classification model based on AADB, AVA, and Flickr-AES datasets [97] are pre-trained. These models are then evaluated by comparing the predicted scores with subjective ratings available for several small-scale fine art datasets. Besides, the global exploratory analysis is performed to explore the relation of the three concepts and high-level art attributes, associated with an analysis of the distribution in different artistic styles or genres.

4) NEW FRAMEWORKS

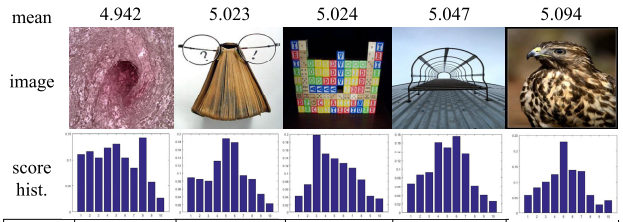
Michal *et al.* [98] investigated the possibility of improving image aesthetic inference of convolutional neural networks with hand-designed features that rely on domain expert feature knowledge in photography. They compare a wide range of handcrafted features to predict binary classification and continuous aesthetic scores and selected 8% and 15% as the best performing feature sets respectively. Moreover, the model achieves excellent performance by

combining the hand-designed features with activation from VGG16 and ResNet50 networks. Li *et al.* [99] presented a personality-assisted multi-task deep learning framework for both generic and personalized image aesthetics assessment, which utilizes an inter-task fusion by training a siamese network to generate individual's personalized aesthetic score on the image. Zhang *et al.* [80] proposed a novel Multi-modal Recurrent Attention Convolutional Neural Network (MRACNN). It consists of two streams: the vision stream and the language stream. The former utilizes the recurrent attention network to eliminate insignificant information and extracted visual features on some important regions. The latter employs the Text-CNN to quantify the high-level semantics of user comments. Finally, a Multimodal Factorized Bilinear pooling approach is used to effectively combine the visual and textual features.

Liu *et al.* [100] proposed a novel semi-supervised deep active learning (SDAL) algorithm, which discovers how humans perceive semantically important regions from a large number of images partially assigned with contaminated tags. Specifically, the SDLA sequentially links semantically important object regions from each scenery and unified these patches with the deeply learned human gaze shifting path (GSP) features into a principled probabilistic model for image aesthetic assessment. Sheng *et al.* [81] proposed an effective self-supervised learning scheme to extract useful feature representation for image aesthetic assessment without manual annotations. Based on the motivation that a suitable feature representation space could identify the changes in aesthetic quality caused by different image editing operations, they proposed two different self-supervised pretext tasks on distinguishing the manipulation types such as downsampling and upsampling, JPEG compression, Gaussian noise, Gaussian blur, color quantization, random patch shuffle, etc, and strength of image degradation operations. The experiments prove that their self-supervised aesthetic-aware feature learning outperforms other self-supervised frameworks in three aesthetics benchmarks.

B. AESTHETIC DISTRIBUTION

The aesthetic scoring including binary classification and regression approaches could not completely describe the aesthetic diversity and subjectivity among users' preferences, the varieties of opinions could be reflected by the probability distribution of image aesthetic scores, as shown in Figure 17. NIMA [82] was proposed to predict the distribution of human opinion ratings in image aesthetic assessment for a given image using a CNN-based model. It is realized by replacing the last layer of the baseline CNN with a fully connected layer with an output of 10 neurons representing the histogram distribution of the 10-level scores, followed by soft-max activations. Then the loss function is modified as the Earth Mover's Distance to measure the distance between the network output and the ground truth histogram distribution of human ratings expressed as an empirical probability mass function, which were annotated in the AVA dataset. Baseline network weights



var.	6.665	5.291	5.161	4.987	4.911
median	4.850	5.012	4.783	5.003	5.076
skew.	0.094	0.003	0.320	0.048	0.134
kur.	1.890	2.368	2.305	2.414	2.613

FIGURE 17. The rating distributions are approximated by the score histograms (1-10) with similar mean scores of images around 5. The hist., var., skew. and kur. are short for histogram, variance, skewness and kurtosis [83].

were initialized by training on ImageNet dataset, and then an end-to-end fine-tuning is performed on high-level aesthetics and low-level technical qualities respectively. Jin *et al.* [83] predicted the aesthetic score distribution of human ratings using deep CNN based on the Cumulative distribution loss function with Jensen-Shannon divergence (CJS-CNN), with a new reliability-sensitive learning method based on the kurtosis of the score distribution.

Cui *et al.* [101] characterized the disagreement among users' aesthetic preferences regarding the same image by learning distribution-oriented aesthetic representation, which is developed based on fully convolutional networks with inputs of arbitrary sizes to eliminate the damage of intrinsic aesthetic appeal of images. Besides, they introduce a new deep semantic-aware hybrid network that incorporates the information from object recognition and scene classification to improve the image aesthetics assessment performance. Xu *et al.* [84] explored an efficient context-aware attention-based model to predict the aesthetic score distribution of human subjective opinions. Specifically, an attention module is presented to supply rich contextual dependencies through multi-level aesthetic details and long-range perception, which are realized in hierarchical and spatial context dimensions. Moreover, the loss function based on the Bhattacharyya distance is introduced to calculate the similarity between the network predicted distribution and the human subjective ground truth, as shown in Figure 18.

C. AESTHETIC ATTRIBUTE

Existing evaluation methods do not provide any details on why the photograph is good or bad, or which attributes contribute to the aesthetic feeling of the photograph since these attributes explicitly predict some of the possible cues that a human might perceive to judge an image. There exist probabilistic dependencies among aesthetic assessment and attributes, as shown in Figure 19.

Some works leverage attributes in learning aesthetic features and training classifiers. Based on the AADB dataset, Kong *et al.* [45] proposed a deep CNN to learn photo aesthetic rating problem assisted by the pair-wise relative ranking

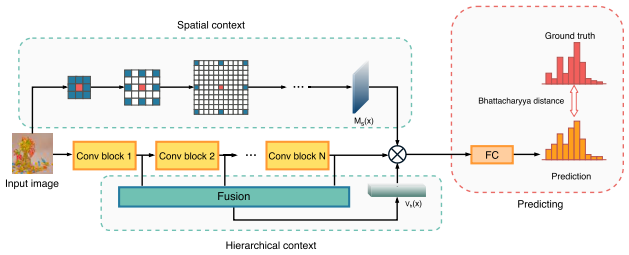
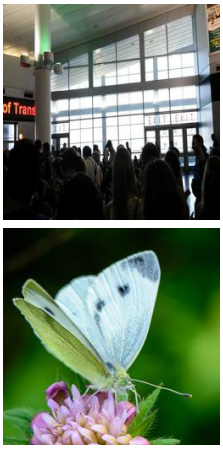


FIGURE 18. The context-aware attention-based framework which predicts the aesthetic score distribution by adopting the Bhattacharyya distance as loss function [84].



Aesthetic score	10%	Motion Blur	50%
Balancing Elements	50%	Object	20%
Color Harmony	30%	Repetition	0%
Content	0%	Rule of Thirds	50%
Depth of Field (DoF)	40%	Symmetry	0%
Light	0%	Vivid Color	10%

Aesthetic score	95%	Motion Blur	50%
Balancing Elements	50%	Object	50%
Color Harmony	90%	Repetition	0%
Content	90%	Rule of Thirds	70%
Depth of Field (DoF)	100%	Symmetry	0%
Light	70%	Vivid Color	100%

FIGURE 19. Two examples of aesthetic images (upper: low aesthetics; lower: high aesthetics) with ratings of the aesthetic score and eleven assessment attributes [87].

modeled in the loss function, and jointly optimized by aesthetic attributes and photo content information. Specifically, they adopt a Siamese architecture that performed pairwise training in the first stage, where the two base networks are pre-trained by fine-tuning AlexNet configurations on aesthetic data using Euclidean Loss regression layer instead of softmax classification layer. Then they utilize a pairwise ranking loss to explicitly exploit relative rankings of image pairs based on the Siamese network as a feature extractor. In the second stage, an attribute-adaptive rating model is trained by adding an attribute prediction branch into the base-net to fuse with the aesthetic branch. In the third stage, they incorporate the content classification branch into the model for joint optimization and predictions, which include predefined category labels. The content branch outputs are used as a weighting vector for gating the combination of predicted scores for aesthetic branch, attribute branch, and content branch. Wang *et al.* [85] designed neuron-inspired deep Chatterjee's Machine (DCM) to learn attributes through the parallel supervised pathways (Figure 20(a)), and then a high-level synthesis network is trained to transform those attributes into the overall aesthetics rating. Although they employ attributes as middle-level representations, the attributes are typically first predicted, and then the predicted attributes are used to measure aesthetics, thus propagating the predicted errors of attributes to the evaluated aesthetic.

Viswanatha *et al.* [86] constructed a novel multi-task deep CNN with a merge-layer, which collects pooled features of the convolution maps to jointly learned eight aesthetic attributes along with the overall aesthetic score simultaneously. To understand the internal representation of these attributes in the learned model, they also develop the visualization technique using backpropagation of gradients, which highlights the key regions for the corresponding attributes. Unlike middle-level representation approaches, a multi-task approach could avoid the predicted errors of attributes propagating to aesthetics. However, it fails to model the distributions among attributes and aesthetics.

Pan *et al.* [87] proposed a novel adversarial learning framework to model the joint distributions of aesthetics and attributes, as shown in Figure 20(b)). During training, they use the aesthetics attributes as privileged information to train an attributes-assisted deep convolutional rating network, which learns the aesthetic score and attributes simultaneously with the supervised loss, which minimizes the error between the prediction and the ground truth label. Through backpropagation in multi-task learning, the gradient of the branch of attributes is beneficial for adjusting better feature representations for aesthetic assessment. In order to further capture the correlation between the aesthetic score and attributes, a discriminator is introduced to distinguish the predictions from the real labels and enforce the rating network to generate the reliable predictions which are closer to the

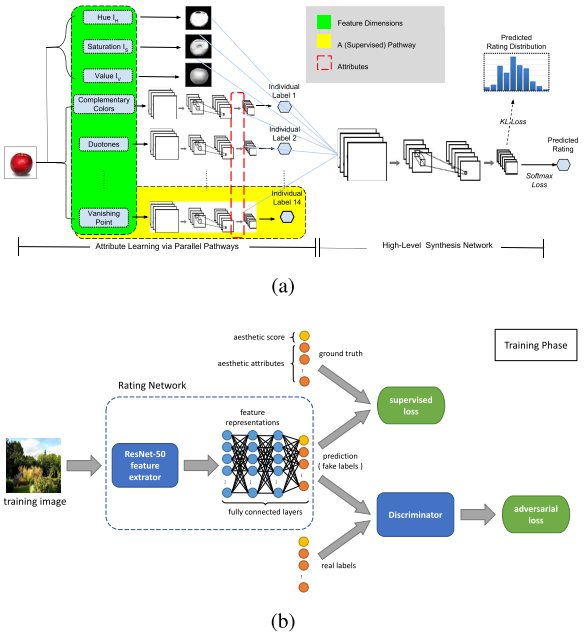


FIGURE 20. Two ways of learning attributes in image aesthetic assessment. (a) attributes are learned as middle-level representations through the parallel supervised pathways [85]. (b) the joint distributions of aesthetic and attributes are modeled by an adversarial learning framework [87].

distribution of the real labels. Under these two optimize objectives, the rating network could efficiently output reliable predictions which minimize the supervised loss and approximate to real distribution of the aesthetic score and attributes simultaneously. Through adversarial learning, the distribution among attributes to aesthetic is fully explored to further regularize aesthetic assessment.

D. AESTHETIC DESCRIPTION

Besides the simple aesthetic quality scoring, it is possible to provide in-depth descriptions and comments in analyzing the reasons why photos are high or low aesthetic appealing in some respect, or why art paintings elicit emotional responses to viewers. The work of Chang *et al.* [46] is the first study that produced captions related to photo aesthetics and/or photography skills. Difference from the common image captioning tasks that depict the objects or their relations in a picture, Chang's captioning approach could generate aesthetic critiques for images with aspect-oriented generated sentences which are more diverse and favorable for humans. Specifically, they proposed two stages to solve the aesthetic critique problem. In their baseline aspect oriented (AO) approach, the training data are divided into disjoint subsets based on the aspects of sentences, and they employ a CNN-LSTM architecture to train the captioning model for every single aspect. Since AO itself could not exploit the interrelated sentences between different aspects to produce a more diverse caption, the aspect fusion (AF) approach is proposed by training the CNN model with a soft-attention layer to predict the aspect-fusion coefficients from the context information, which could leverage the hidden annotations of different aspects and choose the proper combination dynamically over time to generate a more semantically meaningful caption, as shown in Figure 21. The experimental results on the PCCD [46] demonstrate the effectiveness of their approaches for generating aesthetic-oriented captions of images.

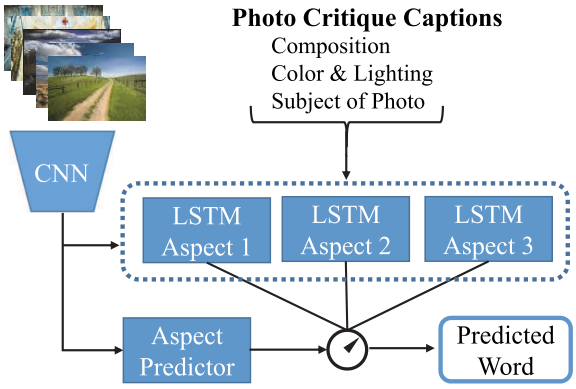


FIGURE 21. The flow diagram of the aspect-oriented (AO) approach [46].

To extend the cognition from rating to reasoning, Wang *et al.* [47] proposed a model referred to as Neural Aesthetic Image Reviewer, which could not only give an aesthetic score for an image, but also generate a textual description explaining the reason why the image is high or low aesthetic. Specifically, they presented CNN plus Recurrent Neural Networks (RNNs) architectures based on shared aesthetically semantic layers and task-specific embedding layers at a high level for performance improvement on different tasks. Through multi-task learning, the proposed model could predict aesthetic scores as well as produce comments in an end-to-end way. The experimental results on the AVA-Reviews dataset [47] verify that the proposed model could generate textual reviews related to aesthetics in consistent with human perception. However, both [46] and [47] can only give a single sentence as the comments describing general image aesthetic impression, which do not describe the individual aesthetic attributes, also the annotations of aesthetic attributes in PCCD are not fully explored.

Jin *et al.* [49] proposed Aesthetic Attributes Assessment of Images, which means the aesthetic attributes captioning. This work is the first attempt to produce both captions and scores for each image aesthetic attribute, including color and lighting, composition, depth and focus, impression and subject, use of camera. By constructing a new dataset named DPC-Captions [49], they proposed Aesthetic Multi-Attribute Network (AMAN), which contains multi-attribute feature network (MAFN), channel and spatial attention network (CSAN), and language generation network (LGN). MAFN measures the feature matrix of 5 attribute scores through the multi-task regression. Due to the fully-annotated small scale of PCCD data, multi-attribute networks are be pre-trained on PCCD and fine-tuned on their weakly-annotated large-scale DPC-Captions. The CSAN dynamically adjusts the attentional weights of channel dimension and spatial dimension of the obtained features. Finally, LGN generates the captions by LSTM network which needs ground truth attribute captions in DPC-Captions and adjusted feature maps from CSAN. The network is evaluated by using both image captioning criteria and mean square error of scoring, which show that the AMAN model outperformed models such as CNN-LSTM in image captions.

Building on the ArtEmis dataset [53] that contains emotional reactions to visual artwork coupled with explanations of these emotions in language, Achlioptas *et al.* [53] developed machine learning models for dominant emotion prediction from images or text by using cross-entropy-based optimization applied to an LSTM text classifier, and trained affective neural speakers that can produce plausible grounded emotion explanations in artworks under three configurations, including baseline with Adjective Noun Pairs (ANPs), basic ArtEmis speakers with two popular backbone architectures including the Show-Attend-Tell approach [102], which combines an image encoder with a word/image attentive LSTM, and the recent line of work of meshed-memory transformers [103], and emotion grounded speaker that promotes the decoupling of the emotion conveyed by the linguistic generation. The experimental results demonstrate that the neural speakers could emulate human emotional responses to visual art and generate associated affective explanations.

V. AESTHETIC-DRIVEN MANIPULATION OF VISUAL ART IMAGES

One of the most common applications in computational aesthetic evaluation is aesthetic-aware image manipulation, the aim of which is using various editing operations to improve the aesthetics of visual art images, as shown in Figure Here we focus on recent literature in three aesthetic enhancement applications including color enhancement, photo recomposition, and aesthetic-guided generation of art paintings.

A. COLOR ENHANCEMENT

1) TRADITIONAL RULE-BASED APPROACHES

Early research works mainly design filter algorithms to enhance the well-established photographic heuristics such as contrast, clarity, exposure, etc. The following three types of approaches are the most representative: the histogram adjustment method [104] dynamically estimates the corresponding mapping function based on local statistical information of the image, thus equalizing the luminance histogram and adjusting it into a specific distribution. The unsharp masking method [105] aims to improve image sharpness. The algorithm decomposes the input image into the base layer and the detail layer, where the weighted scaling factors of different pixels in the detail layer are adjusted adaptively according to the estimated local blur intensity map, and then added back on the base layer to obtain an enhanced version. The retinex-based approaches [106], [107] decompose the photo into reflection and illumination layers for the enhancement of low-light images, to estimate the reflection and piecewise smooth illumination maps under structural responses.

2) DEEP-LEARNING APPROACHES

Due to the strong expressive ability of CNN in learning operations, the deep learning-based color enhancement models have appeared in large numbers with superior results. Ignatov *et al.* [108] constructed a large-scale DSLR photo enhancement dataset (i.e. DPED) consisting of 6K photos taken simultaneously by using a DSLR professional camera and three smartphones, covering various light and weather conditions during the day. Based on these pairwise training data, they propose an end-to-end fully-supervised prediction model to learn a mapping function from low-quality photos taken by mobile phones to high-quality photos taken by DSLR cameras. Ren *et al.* [109] proposed a hybrid network structure for low-light image enhancement. The network consists of two different streams, in which the content flows through estimate the scene content of the low illumination input through the encoder-decoder, and the spatial variation recursive neural network is used as the edge flow to model the image edge details. Chen *et al.* [110] used a scene aggregation model to learn 10 operations commonly used in photo enhancement such as smoothing, restoration, style transfer, fog removal, etc. Park *et al.* [111] regarded the image color enhancement problem as a Markov decision process, and learned the optimal enhancement sequence in each step such as white balance, context, brightness by training agents. In addition, they produced pseudo input and edited image pairs to train the model by performing random color distort manipulation on high-quality reference images. Moran *et al.* [112] trained a deep neural network for regressing the parameters of spatial local filters with paired data. However, the effective performances of these models rely on a large number of carefully aligned pairs of degraded images and corresponding high-quality counterparts.

To avoid dependence on pairwise training data, some works try to use weakly supervised or unsupervised learning to solve the image enhancement problem. Inspired by the image-to-image translation using Generative Adversarial Network (GAN) model [113], Chen *et al.* [114] designed a dual GAN to learn a bidirectional mapping from low-quality source image domain to high-quality target image domain by constraining the cycle consistency loss. Ni *et al.* [115] constructed a unidirectional GAN for unsupervised enhancement, in which the generator consists of codecs embedded with a global attention-based modulation module, and a multi-scale discriminator is used to check the aesthetic value of the generated image, combined with quality loss and fidelity loss to constrain the details. Yang *et al.* [116] recovered image details from rough to fine by using recursive network architecture to train paired data, and then used adversarial learning to train unpaired data for improving the quality of image illumination and color distribution.

To increase the interpretability of the model, Deng *et al.* [117] proposed an enhanced GAN based on weakly supervised learning driven by aesthetic judgment, in which a generator is used to generate a series of enhanced image operator parameters, including piecewise Lab color enhancer, deep filtering-based enhancer, and Image cropping operator with, and a discriminator based on ResNet module is used to measure the aesthetic quality difference between the generated image and the real high-quality image. The works of [118], [119] proposed a GAN-based reinforcement learning model to directly learn the filters in a proper sequence with suitable parameters. They decompose the enhancement process into a series of resolution-independent differentiable filters, such as exposure, contrast, chroma, and gamma correction. Each retouching operation corresponds to a decision-making process in reinforcement learning, and the training is realized by punishing or rewarding the decisions on what action to take next when given the current image state. Based on the work of [82] that uses the NIMA model as the aesthetic quality predictor to effectively tune parameters of image denoising and tone enhancement operators (Figure 22), Du *et al.* [120] designed a progressive image enhancement framework, which generates retouched images in a heuristic process with parameter searching in a group of self-interpretable image filters under the aesthetic guidance of NIMA model.

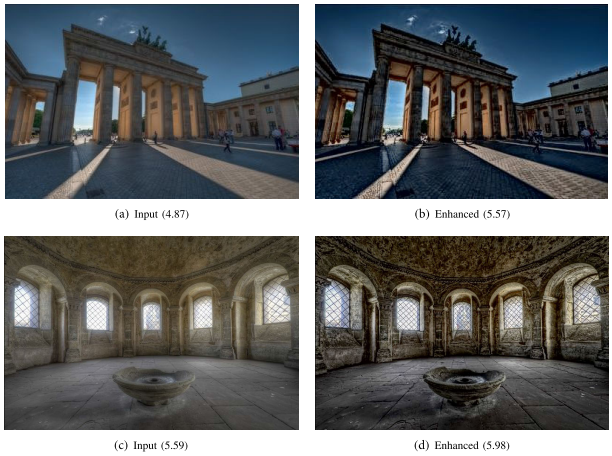


FIGURE 22. Tone enhancement through multi-layer Laplacian technique optimized by the aesthetic assessment model of NIMA with predicted aesthetic scores. (a,c) are the original images, (b,d) are the corresponding enhanced images [82].

B. AESTHETIC-AWARE RECOMPOSITION

Image composition is another manipulation factor that contributes to high aesthetic quality in photos, as shown in Figure 23. Abundant photographic composition rules, e.g., rule of thirds, visual balance are commonly used by professional photographers to capture pleasing photos. In recent years, the direction of aesthetics-driven recomposition computationally adjusts the composition to enhance the aesthetics of an image while preserving semantic context and geometric structures. Here we discuss the state-of-the-art recomposition techniques that can be classified into cropping, discrete rearrangement, warping, and hybrid approaches that utilize a combination of multiple image operators.

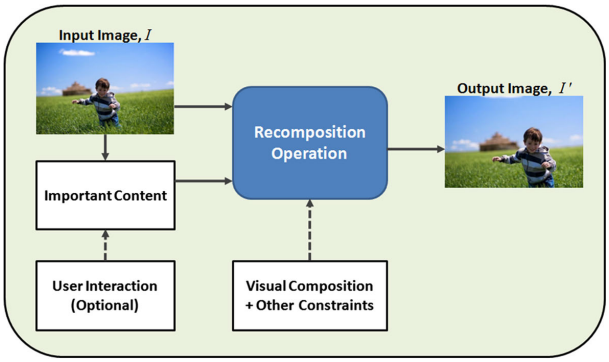


FIGURE 23. The general flow of image recomposition [121].

1) CROPPING

The cropping-based methods search for a cropping window on the original image based on some constraints to retain the significant content to improve the image aesthetics [121]. Previous image cropping schemes can be divided into two aspects. Attention/Saliency-based approaches [122] extract the primary visually salient region in the original image that draws more attention from people, but they may generate unpleasing cropping windows since they ignore the image composition. For aesthetics-based approaches, they aim to find the most pleasing region by evaluating the aesthetics of cropping window candidates based on handcrafted low-level features [123] or deeply-learned CNN features [124]–[126]. The cropping problem is modeled as window candidate classification or regression (e.g. RankNet [127]) used to grade the aesthetic score of the cropping region in a fully supervised learning scheme, which still relies on a limited amount of labeled cropping data and the sliding window method to obtain a large number of candidate windows. Recently, some works formulate the automatic image cropping as a sequential decision-making process in adjusting rectangle to find the best cropping window, and propose aesthetics-aware reinforcement learning frameworks (e.g. A3-RL) with reward functions to address this problem [128], [129]. However, the cropping-based approaches could lose important information if single or multiple objects occupy a significant portion of the image [130].

2) DISCRETE REARRANGEMENT

Discrete approaches rearrange the patches to obtain the recomposed image. The cut-and-paste approaches extract foreground objects from the input image and then paste them back into optimal positions based on a set of photographic composition rules and constraints [121], of which the dependence-aware scheme [131] relocates the foreground objects together with their dependent regions to the optimal position, while the exemplar-based approach relocates the photo subjects [132] using the graph-match optimization. Seam carving-based approaches [133] rearrange a given image by iteratively inserting and removing a set of seams containing significant objects on the contrary direction. Due to its discontinuous property, noticeable feature damages are unavoidable for complex images especially with substantial geometrical features [130].

3) WARPING AND H YBRI

The continuous warping approach recomposed the given image by minimizing a set of aesthetic quality errors measured by popular photographic rules or constraints [121]. Due to its over-compression, the obvious feature distortions are unavoidable in extreme object relocation cases [130]. The hybrid approaches [121] utilize more than one image operator to perform image recomposition. For example, crop-and-warping [134] applied an effective cropping operator to crop off the given image and then applied non-homogeneous warping on the cropped image. To avoid the information loss during the cropping, tearable image warping [135] combined the cut-and-paste and non-homogeneous warping operators to enable the change in a spatial foreground-background relationship while preserving scene consistency.

C. AUTOMATIC GENERATION OF ART PAINTINGS

The advancement of computational aesthetic evaluation would further extend human creativity by inspiring artists and

graphic designers, which makes the automatic generation of artworks possible. A 4-level classification in terms of computational power utilized for computer-generated aesthetic forms of visual art is introduced in [136]. At Level 1, the user could select an existing painting software to draw paintings manually. At Level 2, the user provides various attributes and styles, or mathematic formulas as inputs to generate outputs such as fractal arts. Level 3 utilizes knowledge-based heuristic rules to encode the artists styles into computational algorithms, so as to generate paintings with similar styles, such as computer-generated abstract paintings in Kandinsky style [137], or migrates the styles of paintings to photo images to mimic brush strokes and texture patterns, which is called style transfer [138]. Level 4 heuristic encodes aesthetic rules through machine intelligence to automatically generate highly aesthetic visual forms, which is a potential future research direction.

Specifically, Zheng *et al.* [139] presented a layered approach to generate Pollock's drip style paintings, which are modeled from background layer, irregular shape layer, line layer, and water drop layer sequentially. Tao *et al.* [140], Zhang and Yu [137], Xiong and Zhang [141], Lian *et al.* [142] used parameterized approaches to encode basic visual elements, and randomly generate Malevich, Kandinsky, Miro, Picasso's cubism styles of paintings. Satori *et al.* [64] utilized the aesthetic classification model to guide the generation of abstract paintings that elicit an intended emotional response. They selected three Mondrian paintings and randomly changed the distance, amount, and positions of colors, replacing them with the colors from the color palette generated from the MART dataset to assess the elicited feelings.

By using the GAN-based scheme, He *et al.* [138] proposed ChipGAN, the first weakly supervised deep network architecture in style transfer from photo to Chinese ink wash painting, with three essential techniques constraints: voids, brush strokes, and ink wash tone and diffusion. Kotovenko *et al.* [143] proposed a method to stylize images by aesthetically optimizing parameterized brush strokes instead of pixels and further introduce a simple differentiable rendering mechanism, to avoid the problem of unnatural representation in pixel domain stylization. Zhang *et al.* [144] designed a novel system called AI Painting to generate a specific painting with an illustration of drawing process based on users input, including scene content context, aesthetic effect word, and artistic genre. Specifically, they built a dataset by collecting paintings in six different artistic genres. Then the image content is generated by a StackGAN module and transferred into a specific aesthetic effect based on Image Aesthetic Space with a Bimodal Deep Autoencoder with Cross Edges module; finally, it is simulated with specific artistic genre by neural style transfer and brushstroke enhancement. For the illustration, they displayed the oil painting with the outline and colored strokes layer by layer, and ink paintings with different strokes from subjects to background, as shown in Figure 24.

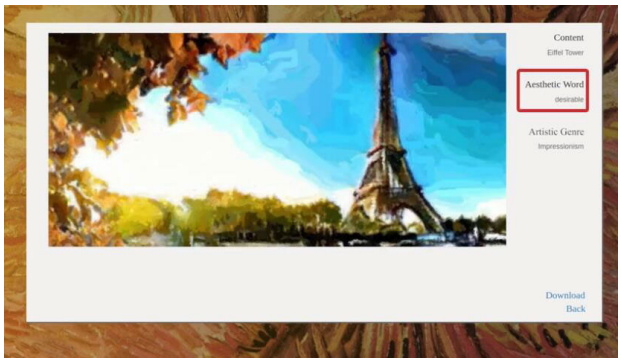


FIGURE 24. The AI painting with illustration of drawing process [144].

VI. OPEN PROBLEMS AND FUTURE DIRECTIONS

In this section, we summarize and discuss several aspects of the potential research problems and valuable future directions. The 4-layer model of aesthetic evaluation of visual art images is shown in Figure 25, in which the level of abstraction, annotation quality, article counts are distributed in a pyramid-like structure: the higher the level of abstraction, the lower quantity, and quality of datasets. The detailed viewpoints are shown as follows.

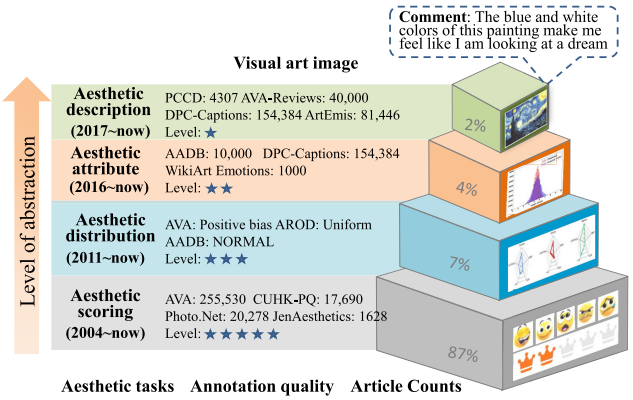


FIGURE 25. The 4 layers of aesthetic evaluation of visual art images.

A. IMPROVEMENT OF HIGH QUALITY DATASETS IN HIGH-LEVEL AESTHETIC TASKS

Compared with image recognition (e.g. ImageNet: 14 million annotated data) and other computer vision tasks, it is difficult to acquire image aesthetic annotations, which the overall scale of datasets are small (e.g. AVA: 255,530). Specifically, current datasets are mainly focused on aesthetic scoring and distribution (more than 300,000), while the number of datasets in aesthetic attribute and description are typically smaller than 100 thousand. The small-scale full annotated PCCD contains both comments and scores of attributes, while the large-scale weakly DPC-Captions only contains partially annotated attribute comments. Besides, in the aesthetic scoring and distribution, each image the typically representative AVA dataset was annotated by at least 78 artists, and the

average rater number is 210. While for the aesthetic attribute and description, the average number of people is less than 20, it is difficult to support the diversity analysis of aesthetic evaluation. It should strengthen the research intensity in the high-level tasks of image aesthetic evaluation, and improve the quality and diversity of annotations in datasets related to the high-level tasks, as well as further extending to aesthetic evaluation research on video quality and graphic designs.

B. BUILDING COMPREHENSIVE AND IN-DEPTH PAINTING AESTHETIC DATASETS

For art paintings, the number of aesthetic assessment datasets (e.g. on average of 1,000) is far less than that in photos, since the public online electronic databases are on a smaller scale, the professional art standards annotation, and restriction of copyright laws in sharing of privately generated collections of artworks. The future work should build more comprehensive and in-depth art painting datasets by considering both quantity, variety, and quality. In the quantity and variety, due to the high-level open access to western oil paintings, the existing painting datasets are mostly focused on western oil or abstract paintings. Along with the continuous advancement of electronic works in an online art museum collection, since there exist significant differences in visual features, semantic features, and aesthetic principles between Chinese and Western paintings with a long history of rich art calendar, the art painting datasets are expected to be gradually completed by enlarging the aesthetic annotations of a large number of Chinese paintings (e.g.OpenSkywork-ChineseClassic Database[1]). In terms of quality, the current aesthetic annotation information in paintings mainly comes from art lovers or machine algorithms, where there may be noise information. Thus it requires inviting art experts in specific domains to check the aesthetic labeling information.

Besides, most of the public painting datasets are concentrated in attribute recognition (e.g. Art500k [145] in 554,000, BAM [146] in 2500,000 with emotion and content descriptions), content understanding (e.g. SemArt [147] in 21,384, Artpedia [148] in 2,930), while the scales of datasets in aesthetic judgment are relatively small, due to the variety of complexity in the data acquisition process. The annotation in attribute recognition mainly consists of title, author, subject, genre, style, etc, and the annotated information of content understanding includes the object class, positions of bounding boxes, and description text, while the information of aesthetic evaluation includes quantitative annotation of aesthetic feeling and emotion experiments should be designed in combination with psychological theories to obtain labeled information. Therefore, it should enlarge the quantity and quality of datasets for high-level aesthetic judgment tasks of art paintings instead of purely aesthetic scoring, such as aesthetic attributes, techniques, emotion and aesthetic appreciation comments in paintings, which could fully utilize the aesthetic knowledge transfer [49] from the rich dataset annotations in attribute recognition and content understanding.

[1]https://openskywork.github.io/OpenSkywork-ChineseClassic

C. INTERPRETABILITY OF DEEP AESTHETIC ASSESSMENT MODELS

At present, the mainstream technique used in image aesthetic evaluation is the deep neural network, which shows outperformed performance than the previous handcrafted aesthetic features. However, the aesthetic learning characteristics of deep neural networks are difficult to be interpreted, which is hard to support the deepen exploration of human aesthetic intelligence. We need to open the black box from deeply-learned aesthetic features in various aesthetic judgment tasks for visual art images. There are several attempts in this field for computer vision tasks such as the Grad-Cam [149] that uses the gradient of the target concept to generate an activation heat map, which is then used to highlight the important pixels for visual interpretation of the decision. The Activation Atlas [150] visualizes feature and attribution by using addition or interpolations between two neurons to demonstrate the semantic arithmetic properties of the activation space and how neurons jointly represent images. For explainable in deep aesthetic models, the predicted aesthetic score for a given attribute could be mapped back to the rectified convolution layers to generate the attribute activation maps, which highlight the attribute-specific discriminative regions [86], or visualizing how the aesthetic classifier ''sees'' paintings while judging the positive or negative emotions by using the back projection technique to display the relative pixel-wise contributions towards the specific task [40], [78], as shown in Figure 26.

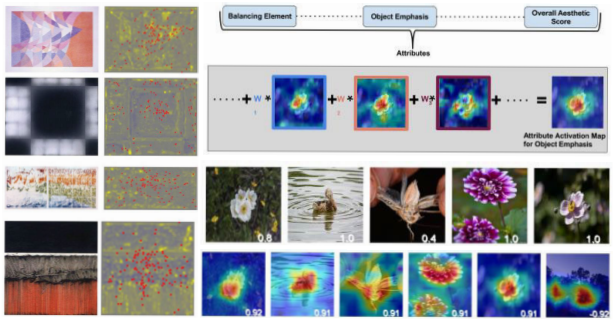


FIGURE 26. Left: Visualizations of pixel-wise contributions to the classification of highly positive emotional (yellow) and negative emotional paintings (blue) [40]. Right: The predicted aesthetic score for a given attribute could be mapped back to the rectified convolution layers to generate the attribute activation maps [86].

D. MULTI-INFORMATION FUSION IN DEEP LEARNING OF ART PAINTINGS

Due to the fewer aesthetic datasets designed for art paintings, currently there exist more researches on applying deep neural networks in painting attribute recognition and content understanding [151], [152] than high-level aesthetic assessment tasks. Sabatelli *et al.* [91] has proved that the deep network

could more focus on the discriminative image regions. Also, the higher degree of correlation with the painting aesthetic evaluation task, the better the model performance of transfer learning is likely to be. Currently, the research on three tasks of attribute recognition, content understanding, and aesthetic judgment, which are three stages in the human aesthetic perception of paintings, are still relatively separated. Therefore, we could utilize the pre-trained deep models in attribute recognition and content understanding to better learn the style and semantic features, which could lead to a better predictive performance in painting aesthetic evaluation.

Besides, the joint utilization of information in attribute recognition, content understanding, and aesthetic evaluation of paintings is another potential direction. We could use the multi-task learning framework to learn multiple painting aesthetic-related attributes simultaneously. By adding more correlation constraints such as artists, styles, object category, scene description as additional supervisory information through the knowledge network, the ability to extract aesthetic information of paintings from the network, and the performance of aesthetic judgment task could be improved. Moreover, since aesthetic appreciation of paintings requires professional background knowledge in the field of art, it should fuse the deep aesthetic model with the hand-designed art professional appreciation features, which is a close interaction between computational aesthetics and art aesthetics.

E. AUTOMATIC DEEP AESTHETICS RECOMPOSITION

Currently, most of the deep learning researches in image aesthetic enhancement has focused on color adjustment and simple cropping. While some effective operations in image recomposition such as cut-and-paste, seam carving, and non-homogeneous warping could only be achieved by manually designed algorithms. It is needed to design photo recomposition-related datasets for training benchmark, and fully utilize the weakly supervised or unsupervised learning to learn a bidirectional mapping from worse-composed to well-composed images. Another line of thought could design a GAN-based reinforcement learning scheme to directly learn the operations in a proper sequence with suitable parameters. Thus the recomposition process could be decomposed into a series of resolution-independent differentiable operations, each retouching operation corresponds to a decision-making process in reinforcement learning, and the training is realized by punishing or rewarding the decisions on what action to take next when given the current image state, which could display the illustration of recomposition process.

F. AESTHETIC-GUIDED ART AESTHETIC ENHANCEMENT

Most studies in aesthetic enhancement have focused on natural photo images. The aesthetic-guided art aesthetic enhancement is also important since it could help beginners to aesthetic evaluation and correction of artworks during the learning and drawing process, which has profound guiding significance in the teaching scenario and art popularization of paintings. Moreover, it has a profound impact on the applications of aesthetic-driven style transfer, computer-aided creation of paintings, and promotional exhibition of digital painting art galleries. However, the existing enhancement models lack the loss constraints designed for aesthetic techniques in Chinese paintings, which may result in undesirable enhancement results. Therefore, we need to explore a suitable constraint model that could simulate the aesthetic techniques of Chinese paintings.

Here are some rational thoughts about this problem. We could use the aesthetic assessment model as the discriminator for feedback guidance, and at the same time, three technique constraints are designed under the weakly-supervised GAN architecture, to achieve the consistency modeling of professional drawing techniques of high-quality Chinese painting images: (1) Whitespace constraint based on adversarial loss. To capture the entropy signal changes caused by the density contrast between whitespace and brushstroke in high-quality Chinese painting, we intend to combine the adversarial loss with the consistency loss of perceived content to ensure that the enhanced image not only has the content of source image, but also converges towards the probability distribution of whitespace in professional Chinese painting. (2) Stroke constraint based on multi-level edge loss. The techniques of brush strokes in Chinese painting emphasize the rich changes of strength, thunders, stretching, and concentration. We plan to use nested edge detector to extract multilevel edge images, which is used to simulate the harmonious distribution of brush strokes in different thicknesses. (3) Ink wash constraint based on filtering loss. We plan to adopt image filter operation for high-quality Chinese painting images and enhanced images to simulate the naturalness of diffusion effects in ink color, and use the filtered image as the discriminator input to calculate the ink wash loss to ensure the tone consistency between enhanced images and professional images.

VII. CONCLUSION

In this work, we systematically review major attempts on aesthetic assessment of two typical types of visual art images, photographs and paintings, with detailed comparisons about the characteristics and shortcomings of the existing methods. To summarize, we have investigated the most commonly used publicly available aesthetic assessment datasets on different categories of art images. Then, conventional approaches based on handcrafted aesthetic features have been reviewed. Besides, we have systematically evaluated recent deep learning techniques that are useful for developing robust models in aesthetic judgment on scoring, distribution, attribute, and description. Further, we have explored an extension of aesthetic evaluation to the applications in aesthetic-driven manipulation including color enhancement, aesthetic-aware recomposition, automatic generation of art paintings through computational approaches. We hope that this survey could serve as a comprehensive reference inspiration for future research on the computational aesthetics in visual media and

its potentially influential applications, which could build a bridge of quantitative aesthetic study between different visual art forms.

REFERENCES

[1] L. Y. Zhu, *Aesthetic Dictionary*. Shanghai, China: Shanghai Lexicographical Publishing House, 2010.
[2] G. T. Fechner, *Vorschule der Aesthetik*, vol. 1. Wiesbaden, Germany: Breitkopf Hartel, 1876.
[3] H. Kawabata and S. Zeki, "Neural correlates of beauty," *J. Neurophysiol.*, vol. 91, no. 4, pp. 1699–1705, 2004.
[4] S. Zeki, "Clive Bell's 'Significant Form' and the neurobiology of aesthetics," *Frontiers Hum. Neurosci.*, vol. 7, p. 730, Dec. 2013.
[5] T. Ishizu and S. Zeki, "The brain's specialized systems for aesthetic and perceptual judgment," *Eur. J. Neurosci.*, vol. 37, no. 9, pp. 1413–1420, May 2013.
[6] S. Brown, X. Gao, L. Tisdelle, S. B. Eickhoff, and M. Liotti, "Naturalizing aesthetics: Brain areas for aesthetic appraisal across sensory modalities," *NeuroImage*, vol. 58, no. 1, pp. 250–258, Sep. 2011.
[7] A. Chatterjee, "Prospects for a cognitive neuroscience of visual aesthetics," *Bull. Psychol. Arts*, vol. 4, no. 2, pp. 56–60, 2004.
[8] B. Wandell, S. Dumoulin, and A. Brewer, "Visual cortex in humans," *Encyclopedia Neurosci.*, vol. 10, pp. 251–257, Dec. 2009.
[9] P. Cavanagh, "The artist as neuroscientist," *Nature*, vol. 434, no. 7031, pp. 301–307, Mar. 2005.
[10] M. Freeman, *The Photographer's Eye: Composition and Design for Better Digital Photos*. Boca Raton, FL, USA: CRC Press, 2007.
[11] H. Leder, B. Belke, A. Oeberst, and D. Augustin, "A model of aesthetic appreciation and aesthetic judgments," *Brit. J. Psychol.*, vol. 95, no. 4, pp. 489–508, Nov. 2004.
[12] F. Hoenig, "Defining computational aesthetics," in *Proc. Eur. Conf. Comput. Aesthet.* Cham, Switzerland: Eurographics Association Aire-la-Ville, 2005, pp. 13–18.
[13] P. A. Fishwick, *Aesthetic Computing*. Cambridge, MA, USA: MIT Press, 2008.
[14] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2006, pp. 288–301.
[15] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, WaShington, DC, USA, Dec. 2006, pp. 419–426.
[16] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2008, pp. 386–399.
[17] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. CVPR*, WaShington, DC, USA, Jun. 2011, pp. 1657–1664.
[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Proces. Syst.*, 2012, pp. 1097–C1105.
[19] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Proces. Syst.*, 2014, pp. 3320–3328.
[20] J. Donahue, Y. G. Jia, and O. Vinyals, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn. (ICML)*. New York, NY, USA, 2014, pp. 32647–32655.
[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[22] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *Proc. ACM Int. Conf. Multimedia (MM)*, New York, NY, USA, 2014, pp. 457–466.
[23] 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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Washington, DC, USA, Dec. 2015, pp. 990–998.
[24] S. Ma, J. Liu, and C. W. Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2017, pp. 4535–4544.
[25] X. Zhang, X. Gao, W. Lu, and L. He, "A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2815–2826, Nov. 2019.
[26] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Process. Mag.*, vol. 28, no. 5, pp. 94–115, Sep. 2011.
[27] G. Zhai and X. Min, "Perceptual image quality assessment: A survey," *Sci. China Inf. Sci.*, vol. 63, no. 11, pp. 211–301, Nov. 2020.
[28] Y. Niu, Y. Zhong, W. Guo, Y. Shi, and P. Chen, "2D and 3D image quality assessment: A survey of metrics and challenges," *IEEE Access*, vol. 7, pp. 782–801, 2019.
[29] A. Raj, A. K. Tiwari, and M. G. Martini, "Fundus image quality assessment: Survey, challenges, and future scope," *IET Image Process.*, vol. 13, no. 8, pp. 1211–1224, Jun. 2019.
[30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[31] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.
[32] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Dec. 2006.
[33] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 80–106, Jul. 2017.
[34] R. Y. Bai, X. Y. Guo, C. H. Jia, and H. J. Geng, "Overview of research methods of painting aesthetics," *J. Image Graph.*, vol. 24, no. 11, pp. 1860–1881, 2019.
[35] M. Fiorucci, M. Khoroshiltseva, M. Pontil, A. Traviglia, A. Del Bue, and S. James, "Machine learning for cultural heritage: A survey," *Pattern Recognit. Lett.*, vol. 133, pp. 102–108, May 2020.
[36] M. Perc, "Beauty in artistic expressions through the eyes of networks and physics," *J. Roy. Soc. Interface*, vol. 17, no. 164, Mar. 2020, Art. no. 20190686.
[37] Y. Lu, C. Guo, Y. L. Lin, F. Zhuo, and F. Y. Wang, "Computational aesthetics of fine art paintings: The state of the art and outlook," *Acta Automatica Sinica*, vol. 46, no. 11, pp. 2239–2259, 2020.
[38] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1930–1943, Dec. 2013.
[39] C. Li and T. Chen, "Aesthetic visual quality assessment of paintings," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 236–252, Apr. 2009.
[40] A. Sartori, V. Yanulevskaya, A. A. Salah, J. Uijlings, E. Bruni, and N. Sebe, "Affective analysis of professional and amateur abstract paintings using statistical analysis and art theory," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 2, pp. 1–27, Jul. 2015.
[41] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2408–2415.
[42] A. Agrawal, V. Premachandran, and R. Kakarala, "Rating image aesthetics using a crowd sourcing approach," in *Proc. 6th Pacific-Rim Symp. Image Video Technol.*, 2013, pp. 24–32.
[43] R. Schifanella, M. Redi, and L. M. Aiello, "An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures," in *Proc. 9th AAAI Int. Conf. Weblogs Social Media*, 2015, pp. 296–317.
[44] K. Schwarz, P. Wieschollek, and H. P. A. Lensch, "Will people like your image? Learning the aesthetic space," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 2048–2057.
[45] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Feb. 2016, pp. 662–679.
[46] K.-Y. Chang, K.-H. Lu, and C.-S. Chen, "Aesthetic critiques generation for photos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3514–3523.
[47] W. Wang, S. Yang, W. Zhang, and J. Zhang, "Neural aesthetic image reviewer," *IET Comput. Vis.*, vol. 13, no. 8, pp. 749–758, Dec. 2019.
[48] Y. Zhou, X. Lu, J. Zhang, and J. Z. Wang, "Joint image and text representation for aesthetics analysis," in *Proc. ACM Conf. Multimedia Conf.*, 2016, pp. 262–266.
[49] X. Jin, L. Wu, G. Zhao, X. Li, X. Zhang, S. Ge, D. Zou, B. Zhou, and X. Zhou, "Aesthetic attributes assessment of images," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 311–319.
[50] S. A. Amirshahi, G. U. Hayn-Leichsenring, J. Denzler, and C. Redies, "Jenaesthetics subjective dataset: Analyzing paintings by subjective scores," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2015, pp. 3–19.

[51] Y. Zhan, Y. Gao, and L. Y. Xie, "Feature analysis and classification for aesthetic of chinese traditional painting," *J. Beijing Univ. Aeronaut. Astronaut.*, vol. 45, no. 12, pp. 2514–2522, 2019.

[52] S. Mohammad and S. Kiritchenko, "Wikiart emotions: An annotated dataset of emotions evoked by art," in *Proc. 11th Int. Conf. Lang. Resour. Eval.*, 2018, pp. 1225–1238.

[53] P. Achlioptas, M. Ovsjanikov, K. Haydarov, and, "Artemis: Affective language for visual art," 2019, *arXiv:1312.6229*. [Online]. Available: https://arxiv.org/abs/2101.07396

[54] B. Thomee, D. A. Shamma, G. Friedland, and, "The new data and new challenges in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2015.

[55] Y. Zhan, Y. Gao, and L. Y. Xie, "Database of emotion and aesthetics in chinese paintings," *J. Image Graph.*, vol. 12, pp. 2267–2278, Dec. 2019.

[56] X. Sun, H. Yao, R. Ji, and S. Liu, "Photo assessment based on computational visual attention model," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 541–544.

[57] T. O. Aydin, A. Smolic, and M. Gross, "Automated aesthetic analysis of photographic images," *IEEE Trans. Vis. Comput. Graphics*, vol. 21, no. 1, pp. 31–42, Jan. 2015.

[58] Y. S. Rawat and M. S. Kankanhalli, "Context-aware photography learning for smart mobile devices," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 12, no. 1s, p. 19, 2015.

[59] A. Lienhard, P. Ladret, and A. Caplier, "Low level features for quality assessment of facial images," in *Proc. 10th Int. Conf. Comput. Vis. Theory Appl.*, 2015, pp. 545–552.

[60] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1784–1791.

[61] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *Proc. CVPR*, Jun. 2011, pp. 33–40.

[62] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien, "Preference-aware view recommendation system for scenic photos based on Bag-of-Aesthetics-Preserving features," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 833–843, Jun. 2012.

[63] B. Mallon, C. Redies, and G. U. Hayn-Leichsenring, "Beauty in abstract paintings: Perceptual contrast and statistical properties," *Frontiers Hum. Neurosci.*, vol. 8, pp. 1–14, Mar. 2014.

[64] A. Sartori, D. Culibrk, Y. Yan, R. Job, and N. Sebe, "Computational modeling of affective qualities of abstract paintings," *IEEE MultimediaMag.*, vol. 23, no. 3, pp. 44–54, Feb. 2016.

[65] J. J. Zhang, R. Peng, J. Wang, and J. H. Yu, "Computational aesthetic evaluation of chinese wash paintings," *J. Softw.*, vol. 27, no. S2, pp. 220–233, 2017.

[66] L.-K. Wong and K.-L. Low, "Saliency-enhanced image aesthetics class prediction," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 997–1000.

[67] C. C. Li, A. C. Loui, and T. Chen, "Towards aesthetics: A photo quality assessment and photo selection system," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 827–830.

[68] L. Marchesotti and F. Perronnin, "Learning beautiful attributes," in *Proc. Brit. Mach. Vis. Conf.*, 2013, pp. 1–11.

[69] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2004, pp. 1–2.

[70] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2007, pp. 1–8.

[71] A. Sartori, D. Culibrk, Y. Yan, and N. Sebe, "Who's afraid of itten: Using the art theory of color combination to analyze emotions in abstract paintings," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 311–320.

[72] A. Sartori, Y. Yan, G. Ozbal, A. A. A. Salah, and N. Sebe, "Looking at mondrian's victory boogie-woogie: What do i feel?" in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 2503–2509.

[73] S. A. Amirshahi and J. Denzler, "Judging aesthetic quality in paintings based on artistic inspired color features," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl.*, Mar. 2017, pp. 1–8.

[74] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 497–506.

[75] W. Wang, M. Zhao, L. Wang, J. Huang, C. Cai, and X. Xu, "A multi-scene deep learning model for image aesthetic evaluation," *Signal Process., Image Commun.*, vol. 47, pp. 511–518, Sep. 2016.

[76] Y. Kao, R. He, and K. Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1482–1495, Mar. 2017.

[77] C. Li, S. Q. Sun, X. Min, W. X. Wang, and Z. C. Tang, "Application of deep convolutional features in sketch works classification and evaluation," *J. Comput. Aided Des. Comput. Graph.*, vol. 27, no. 10, pp. 1898–1904, 2017.

[78] H. Zhang and D. Xu, "Ethnic painting analysis based on deep learning," *Scientia Sinica Inf.*, vol. 49, no. 2, pp. 204–215, Feb. 2019.

[79] J. Zhang, Y. Miao, J. Zhang, and J. Yu, "Inkthetics: A comprehensive computational model for aesthetic evaluation of chinese ink paintings," *IEEE Access*, vol. 8, pp. 225857–225871, 2020.

[80] X. Zhang, X. Gao, W. Lu, L. He, and J. Li, "Beyond vision: A multimodal recurrent attention convolutional neural network for unified image aesthetic prediction tasks," *IEEE Trans. Multimedia*, vol. 23, pp. 611–623, 2021.

[81] K. Sheng, W. Dong, M. Chai, and, "Revisiting image aesthetic assessment via self-supervised feature learning," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 5709–5716.

[82] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[83] X. Jin, L. Wu, X. Li, and, "Predicting aesthetic score distribution through cumulative Jensen-Shannon divergence," in *Proc. 32th AAAI Conf. Artif. Intell.*, 2018, pp. 77–84.

[84] M. N. Xu, J. X. Zhong, Y. R. Ren, and, "Context-aware attention network for predicting image aesthetic subjectivity," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 798–805.

[85] Z. Wang, D. Liu, S. Chang, F. Dolcos, D. Beck, and T. Huang, "Image aesthetics assessment using deep Chatterjee's machine," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 941–948.

[86] G. Viswanatha Reddy, S. Mukherjee, and M. Thakur, "Measuring photography aesthetics with deep CNNs," *IET Image Process.*, vol. 14, no. 8, pp. 1561–1570, Jun. 2020.

[87] B. Pan, S. Wang, and Q. Jiang, "Image aesthetic assessment assisted by attributes through adversarial learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 21, 2019, pp. 679–686.

[88] Y. Chen, Y. Hu, L. Zhang, P. Li, and C. Zhang, "Engineering deep representations for modeling aesthetic perception," 2016, *arXiv:1605.07699*. [Online]. Available: http://arxiv.org/abs/1605.07699

[89] K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang, and B.-G. Hu, "Attention-based multi-patch aggregation for image aesthetic assessment," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1–8.

[90] X. Tian, Z. Dong, K. Yang, and T. Mei, "Query-dependent aesthetic model with deep learning for photo quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2035–2048, Nov. 2015.

[91] M. Sabatelli, M. K. W. Daelemans, and P. Geurts, "Deep transfer learning for art classification problems," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 631–646.

[92] M. Bojarski, A. Choromanska, K. Choromanski, and, "Visualbackprop: Efficient visualization of CNNs," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Mar. 2018, pp. 4701–4708.

[93] T. Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Mar. 2015, pp. 1449–1457.

[94] E. Cetinic, T. Lipic, and S. Grgic, "A deep learning perspective on beauty, sentiment, and remembrance of art," *IEEE Access*, vol. 7, pp. 73694–73710, 2019.

[95] Q. Z. You, J. B. Luo, H. L. Jin, and J. C. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 381–388.

[96] M. Katsurai and S. Satoh, "Image sentiment analysis using latent correlations among visual, textual, and sentiment views," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2837–2841.

[97] J. Ren, X. H. Shen, Z. Lin, and, "Personalized image aesthetics," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 638–647.

[98] K. Michal, C. L. Alexander, and W. M. David, "Leveraging expert feature knowledge for predicting image aesthetics," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5100–5122, Dec. 2018.

[99] L. Li, H. Zhu, S. Zhao, G. Ding, and W. Lin, "Personality-assisted multi-task learning for generic and personalized image aesthetics assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 3898–3910, 2020.

[100] Z. Liu, Z. Wang, Y. Yao, L. Zhang, and L. Shao, "Deep active learning with contaminated tags for image aesthetics assessment," *IEEE Trans. Image Process.*, early access, Apr. 18, 2018, doi: 10.1109/TIP.2018.2828326.
[101] C. Cui, H. Liu, T. Lian, L. Nie, L. Zhu, and Y. Yin, "Distribution-oriented aesthetics assessment with semantic-aware hybrid network," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1209–1220, May 2019.
[102] K. Xu, J. Ba, R. Kiros, and, "Show, attend and tell: Neural image caption generation with visual attention," in *Int. Conf. Mach. Learn. (ICML)*, Jan. 2015, pp. 2048–2057.
[103] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, p. 10.
[104] C. Lee, C. Lee, and C.-S. Kim, "Contrast enhancement based on layered difference representation of 2D histograms," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5372–5384, Dec. 2013.
[105] W. Ye and K. K. Ma, "Blurriness-guided unsharp masking," *IEEE Trans. Image Process*, vol. 27, no. 9, pp. 4465–4477, 2018.
[106] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust retinex model," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2828–2841, Jun. 2018.
[107] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, Feb. 2017.
[108] A. Ignatov, N. Kobyshev, R. Timofte, and K. Vanhoey, "DSLR-quality photos on mobile devices with deep convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3277–3285.
[109] W. Ren, S. Liu, L. Ma, Q. Xu, X. Xu, X. Cao, J. Du, and M.-H. Yang, "Low-light image enhancement via a deep hybrid network," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4364–4375, Sep. 2019.
[110] Q. Chen, J. Xu, and V. Koltun, "Fast image processing with fully-convolutional networks," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2516–2525.
[111] J. Park, J. Lee, D. Yoo, and, "Distort-and-recover: Color enhancement using deep reinforcement learning," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 5928–5936.
[112] S. Moran, P. Marza, S. McDonagh, S. Parisot, and G. Slabaugh, "DeepLPF: Deep local parametric filters for image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, p. 12.
[113] J. Y. Zhu, T. Park, P. Isola, and, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput*, Mar. 2017, pp. 2223–2232.
[114] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6306–6314.
[115] Z. Ni, W. Yang, S. Wang, L. Ma, and S. Kwong, "Towards unsupervised deep image enhancement with generative adversarial network," *IEEE Trans. Image Process.*, vol. 29, pp. 9140–9151, 2020.
[116] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3060–3069.
[117] Y. Deng, C. C. Loy, and X. Tang, "Aesthetic-driven image enhancement by adversarial learning," in *Proc. ACM Conf. Multimedia*, 2018, pp. 870–878.
[118] Y. Hu, H. He, C. Xu, B. Wang, and S. Lin, "Exposure: A white-box photo post-processing framework," *ACM Trans. Graph.*, vol. 37, no. 2, p. 26, 2018.
[119] S. Kosugi and T. Yamasaki, "Unpaired image enhancement featuring reinforcement-learning-controlled image editing software," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 34, no. 7, pp. 11296–11303.
[120] X. Du, X. Yang, Z. Qin, and J. Tang, "Progressive image enhancement under aesthetic guidance," in *Proc. ACM SIGMM Int. Conf. Multimedia Retr. (ICMR)*, 2019, pp. 349–353.
[121] M. B. Islam, W. Lai-Kuan, and W. Chee-Onn, "A survey of aesthetics-driven image recomposition," *Multimedia Tools Appl.*, vol. 76, no. 7, pp. 9517–9542, Apr. 2017.
[122] C. Fang, Z. Lin, R. Mech, and X. Shen, "Automatic image cropping using visual composition, boundary simplicity and content preservation models," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1105–1108.
[123] J. Yan, S. Lin, S. B. Kang, and X. Tang, "Learning the change for automatic image cropping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2013, pp. 971–978.
[124] G. Guo, H. Wang, C. Shen, Y. Yan, and H.-Y.-M. Liao, "Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2073–2085, Aug. 2018.
[125] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1531–1544, Jul. 2019.
[126] H. Zeng, L. Li, Z. Cao, and L. Zhang, "Reliable and efficient image cropping: A grid anchor based approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5949–5957.
[127] Z. Wei, J. Zhang, X. Shen, Z. Lin, R. Mech, M. Hoai, and D. Samaras, "Good view hunting: Learning photo composition from dense view pairs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5437–5446.
[128] D. Li, H. Wu, J. Zhang, and K. Huang, "Fast A3RL: Aesthetics-aware adversarial reinforcement learning for image cropping," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5105–5120, Oct. 2019.
[129] X. Zhang, Z. Li, and J. Jiang, "Emotion attention-aware collaborative deep reinforcement learning for image cropping," *IEEE Trans. Multimedia*, early access, Aug. 4, 2021, doi: 10.1109/TMM.2020.3013350.
[130] M. B. Islam, L.-K. Wong, K.-L. Low, and C.-O. Wong, "Aesthetics-driven stereoscopic 3-D image recomposition with depth adaptation," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 2964–2979, Nov. 2018.
[131] F. L. Zhang, M. Wang, and S. M. Hu, "Aesthetics-driven stereoscopic 3-D image recomposition with depth adaptation," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1480–C1490, Oct. 2013.
[132] H.-T. Chang, Y.-C.-F. Wang, and M.-S. Chen, "Transfer in photography composition," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, p. 957.
[133] K. Li, B. Yan, J. Li, and A. Majumder, "Seam carving based aesthetics enhancement for photos," *Signal Process., Image Commun.*, vol. 39, pp. 509–516, Nov. 2015.
[134] Y. Jin, Q. Wu, and L. Liu, "Aesthetic photo composition by optimal crop-and-warp," *Comput. Graph.*, vol. 36, no. 8, pp. 955–965, 2012.
[135] L. K. Wong and K. L. Low, "Tearable image warping for extreme image retargeting," in *Proc. 30th Comput. Graph. Int. Conf.*, 2012, pp. 1–8.
[136] K. Zhang, S. Harrell, and X. Ji, "Computational aesthetics: On the complexity of computer-generated paintings," *Leonardo*, vol. 45, no. 3, pp. 243–248, Jun. 2012.
[137] K. Zhang and J. Yu, "Generation of kandinsky art," *Leonardo*, vol. 49, no. 1, pp. 48–54, Feb. 2016.
[138] B. He, F. Gao, D. Ma, B. Shi, and L.-Y. Duan, "ChipGAN: A generative adversarial network for chinese ink wash painting style transfer," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1172–1180.
[139] Y. Zheng, X. Nie, Z. Meng, W. Feng, and K. Zhang, "Layered modeling and generation of Pollock's drip style," *Vis. Comput.*, vol. 31, no. 5, pp. 589–600, May 2015.
[140] W. Tao, Y. Liu, and K. Zhang, "Automatically generating abstract paintings in malevich style," in *Proc. IEEE/ACIS 13th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2014, pp. 201–205.
[141] L. Xiong and K. Zhang, "Generation of miro's surrealism," in *Proc. 9th Symp. Vis. Inf. Commun. Interact. (VINCI)*, 2016, pp. 130–137.
[142] G. Y. Lian, Y. L. Wang, K. Zhang, and L. Yao, "An attempt in modeling picasso's cubism style," in *Proc. Symp. Vis. Inf. Commun. Interact. (VINCI)*, 2016, pp. 1–2.
[143] D. Kotovenko, M. Wright, A. Heimbrecht, and, "Rethinking style transfer: From pixels to parameterized brushstrokes," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2021, pp. 1–5.
[144] C. J. Zhang, K. H. Lei, J. Jia, and, "Ai painting: An aesthetic painting generation system," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 1231–1233.
[145] H. Mao, M. Cheung, and J. She, "Deepart: Learning joint representations of visual arts," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 1183–1191.
[146] M. J. Wilber, C. Fang, H. L. Jin, and, "BAM! The behance artistic media dataset for recognition beyond photography," in *Proc. IEEE Int. Conf. Comput. Vis.*, Feb. 2017, pp. 1202–1211.
[147] N. Garcia and G. Vogiatzis, "How to read paintings: Semantic art understanding with multi-modal retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 676–691.

[148] M. Stefanini, M. Cornia, L. Baraldi, and, "Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain," in *Proc. 20th Int. Conf. Image Anal. Process.*, 2019, pp. 729–740.
[149] R. R. Selvaraju, M. Cogswell, and A. Das, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.
[150] *Exploring Neural Networks With Activation Atlases*. Accessed: Jun. 12, 2020. [Online]. Available: https://distill.pub/2019/activation-atlas/
[151] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5001–5009.
[152] S. R. Sheng and M. F. Moens, "Deep transfer learning for art classification problems," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 2478–2486.

JIAJING ZHANG was born in Hebei, China, in 1991. She received the B.S. degree in computer science and technology from the South China University of Technology, Guangzhou, Guangdong, China, in 2012, and the Ph.D. degree in computer science and technology from Zhejiang University, Hangzhou, Zhejiang, China, in 2017.

Since 2018, she has been a Lecturer with the Information Science and Technology Department, Zhejiang Sci-Tech University, Hangzhou. Her current research interests include image computational aesthetic assessment, visual media computing, image processing, and deep learning.

YONGWEI MIAO (Member, IEEE) received the master's degree in mathematics from the Institute of Mathematics, Chinese Academy of Sciences, Beijing, in July 1996, and the Ph.D. degree in computer graphics from the State Key Laboratory of Computer-aided Design and Computer Graphics, Zhejiang University, Hangzhou, Zhejiang, in March 2007. He worked as a Visiting Scholar with the University of Zurich, Switzerland, from February 2008 to February 2009, and the University of Maryland, USA, from November 2011 to May 2012. From July 2015 to August 2015, he worked as a Visiting Professor with the University of Tokyo, Japan. He is currently a Professor with the College of Information Science and Technology, Zhejiang Sci-Tech University, China. He is the author or coauthor of more than 130 technical articles published in scientific journals or presented at conferences. His research interests include computer graphics, 3-D computer vision, 3-D reconstruction, visual media computing, and deep learning.

JINHUI YU received the B.S. and M.S. degrees from the Department of Electronic Engineering, Harbin Institute of Marine Engineering, Heilongjiang, in 1987, and the Ph.D. degree from the Department of Computer Science and Technology, University of Glasgow, U.K., in 1999.

Since 2001, he has been a Professor with the State Key Laboratory of Computer-aided Design and Computer Graphics, and a Doctoral Supervisor of computer application technology and digital art with Zhejiang University, Hangzhou, Zhejiang. His research interests include computational aesthetics of traditional chinese art, non-photorealistic rendering, and computer animation.

Prof. Yu is a member of ACM SIGGRAPH and the Professional Committee Member of computer animation and digital entertainment with China Image Graphic Society, Digital Art, and China Animation Society.