IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 August 2022, accepted 16 September 2022, date of publication 26 September 2022,
date of current version 30 September 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3209196*

SURVEY

# Image Aesthetic Assessment: A Comparative Study of Hand-Crafted & Deep Learning Models

**ABBAS ANWAR[1], SAIRA KANWAL[2], MUHAMMAD TAHIR[3], (Senior Member, IEEE), MUHAMMAD SAQIB[4,5], MUHAMMAD UZAIR[2], MOHAMMAD KHALID IMAM RAHMANI[3], (Senior Member, IEEE), AND HABIB ULLAH[6]**

[1]Department of Computer Science, Abdul Wali Khan University Mardan (AWKUM), Mardan, Khyber Pakhtunkhwa 23200, Pakistan
[2]Department of Electrical Engineering, COMSATS University Islamabad, Wah Campus, Wah Cantt 47040, Pakistan
[3]College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia
[4]Imaging and Computer Vision Group, Data61-CSIRO, Broadway, NSW 2007, Australia
[5]School of Computer Science, Faculty of Engineering and IT, University of Technology Sydney, Sydney, NSW 2007, Australia
[6]Faculty of Science and Technology, Norwegian University of Life Sciences, 1430 Ås, Norway

Corresponding author: Muhammad Tahir (m.tahir@seu.edu.sa)

**ABSTRACT** Automatic image aesthetics assessment is a computer vision problem dealing with categorizing images into different aesthetic levels. The categorization is usually done by analyzing an input image and computing some measure of the degree to which the image adheres to the fundamental principles of photography such as balance, rhythm, harmony, contrast, unity, look, feel, tone, and texture. Due to its diverse applications in many areas, automatic image aesthetic assessment has gained significant research attention in recent years. This article presents a comparative study of different automatic image aesthetics assessment techniques from the year 2005 to 2021. A number of conventional hand-crafted as well as modern deep learning-based approaches are reviewed and analyzed for their performance on various publicly available datasets. Additionally, critical aspects of different features and models have also been discussed to analyze their performance and limitations in different situations. The comparative analysis reveals that deep learning based approaches excel hand-crafted based techniques in image aesthetic assessment.

**INDEX TERMS** Image aesthetic assessment, aesthetic visual perception, image quality assessment, computer vision, convolutional neural networks, deep learning.

## I. INTRODUCTION

It may be true that beauty lies in the eyes of the beholder but for a computer, automatically quantifying the beauty of a photograph is a challenging task. In computer vision, the task is known as automatic image aesthetics assessment and deals with quantifying the beauty, quality, and impression of photographs to categorize images into different aesthetic levels as shown in Figure 1. Image aesthetic assessment has diverse applications in the field of multimedia content generation and processing including medical & healthcare, information & communication technologies, infotainment, edutainment, and safety & security etc. For example, it can be employed to benchmark the algorithms for image noise removal and image restoration as well as for monitoring of quality of service (QoS) in systems where images are digitally compressed, communicated, and decompressed.

Underwater image enhancement and restoration systems can also benefit from aesthetic assessment techniques [1]. For instance, in underwater systems [2], image quality assessment can be used, and image enhancement approaches can be applied to improve quality and accuracy in case of low-quality image as input [3], [4] [5]. Moreover, image quality assessment can be utilized in robotics, where a robot automatically assesses the image quality and change focus and position to recapture the image if the quality metric is below some recommended level.

Due to its significant application potential in the rapidly increasing digital camera and photography industry, automated image aesthetic assessment has recently gained

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

considerable research attention from the computer vision and pattern recognition community [6], [7], [8], [9], [10]. Automatic image aesthetic assessment has many challenges. For example, the input visual data may contain noise and image artifacts such as illumination and environmental conditions [11]. Focus and pose deflections introduce disparities in images. Images may be subject to variations in colour harmony because of sensor resolution issues. Background clutter can also hinder the accuracy of aesthetic assessment algorithms. Moreover, the visual judgment conflicts of humans also translate to different challenges for image quality rating algorithms.

Over the past couple of decades, many computer vision techniques have been developed for image aesthetic/quality assessment. Both hand-crafted feature-based approaches and deep learning-based approaches have been exploited for the task. Hand-crafted features-based algorithms generally design filters to encode aspects of the image aesthetics such as photographic rules, image texture, local and global content features, etc. The represented aesthetics features are then fed to classical machine learning approaches to classify the image in different aesthetic levels. Deep learning-based techniques use robust deep neural networks to learn and encode image aesthetics from a large number of training images. Deep learning-based methods are more accurate as they can model more complex image features and their relationships.

This article provides a survey of techniques for automatic image aesthetic assessment. Both hand-crafted feature-based methods and the recent deep learning approaches are covered in detail, describing each technique's basic framework with its pros and cons. The outcomes of experimental method in terms of accuracy, the dataset used, its size, and the depth of each aesthetic rating algorithm are also discussed.

*Motivation:* We would like to emphasize that a survey is required in image aesthetics due to many papers published in deep learning; although a review of image aesthetics [4] was published half a decade ago, we argue that the number of articles published in the last year is significantly higher during the previous five years, therefore we expect further increase in the coming years. Furthermore, the review of [4] is more on the lines of explaining the image aesthetics and lacks in listing hand-crafted or deep learning methods in detail. Similarly, we want to provide detailed descriptions of important articles to help the community adopt the most appropriate approach and avoid reproducing the methodologies. Likewise, we strive to give a good research direction through this survey, specify the gaps and limitations, and provide future direction.

## II. HAND-CRAFTED METHODS

Although hand-crafted features are considered a thing of the past, they still provide good insight into a computer vision task. Hand-crafted methods primarily design some kinds of pixel filters to extract or encode low-level image features. Standard features used by the hand-crafted techniques include colour, contrast, saturation, brightness, texture and foreground-background statistics, global features, and local features ratio statistics [36]. Figure 2 summarizes different techniques based on the features they use to encode the aesthetic information about images. We discuss each of these categories in detail in the following sections.

### A. BASIC METHODS

These methods are pioneering image aesthetic approaches and provide a naive methodology for accuracy.

1) An intelligent photographic interface is proposed by Lo *et al.* [13], with on-device aesthetic quality assessment for bi-level image quality on general portable devices (Figure 3(a)). In this framework, photographic rules were followed, and a three-layered structure was designed. Using hand-tuned techniques, the first layer extracted composition, saturation, colour combinations, contrast, and richness features. In the second layer, an independent SVM classifier [37] was trained for each feature perspective to obtain the feature index. Moreover, the SVM classifier is trained to get the aesthetic score in the last layer. The mentioned framework is tested on CUHK [38] dataset, comprising 2078 high-quality and 7573 low-quality images, providing an accuracy of 89%.
2) A computational algorithm using region-based features and k-means clustering is presented by Datta *et al.* [14]. Colour segments are extracted from the image utilizing region-based features and texture information to assess the quality of images with the connected component technique. Subsequently, the SVM classifier on the extracted feature is trained to categorize images into high and low aesthetic categories. A regression model [39] is also trained to obtain a regression score. The dataset is collected from a photo-sharing website consisting of 3581 images.
3) To access the quality of digital portraits, Redi *et al.* [15] introduced a technique based on composition, scene semantics, portrait-specific features, correct perception of signal and fuzzy properties, and the five essential features extracted from images. One should note that composition rules are the essential and basic photography rules, including sharpness, spatial arrangement, lighting, texture, and colour. The semantic contents represent the overall photography depiction, including high-level features [40]. The correct perception of signals includes noise, contrast quality, exposure quality, and JPEG quality, while portrait-specific features include face position, face orientation, age, gender, eye, nose, mouth position, foreground, and background contrast. Fuzzy properties are originality, memorability, uniqueness, and emotion depiction. LASSO regression [41] is applied to the extracted composition features, learning regression parameters for every feature group. Moreover, a correlation between the predicted score and the original aesthetic value is computed. Using regression on all features, a final aesthetic


**FIGURE 1.** The sample images are taken from the Aesthetics and Attributes Database (AADB) [12], consisting of various photographic imagery of real scenes collected from Flickr. For each image, the rating is provided by averaging five rater's score as a ground-truth score. Eleven aesthetic attributes were considered while curating the dataset, such as *interesting content, object emphasis, good lighting, colour harmony, vivid colour, shallow depth of field, motion blur, rule of thirds, balancing element, repetition, and symmetry.* The photos a) and b) represent photos with a high aesthetic score of 1.0 and 0.7, while c) and d) represent the photos with a low aesthetic score of 0.4 and 0.1, respectively.

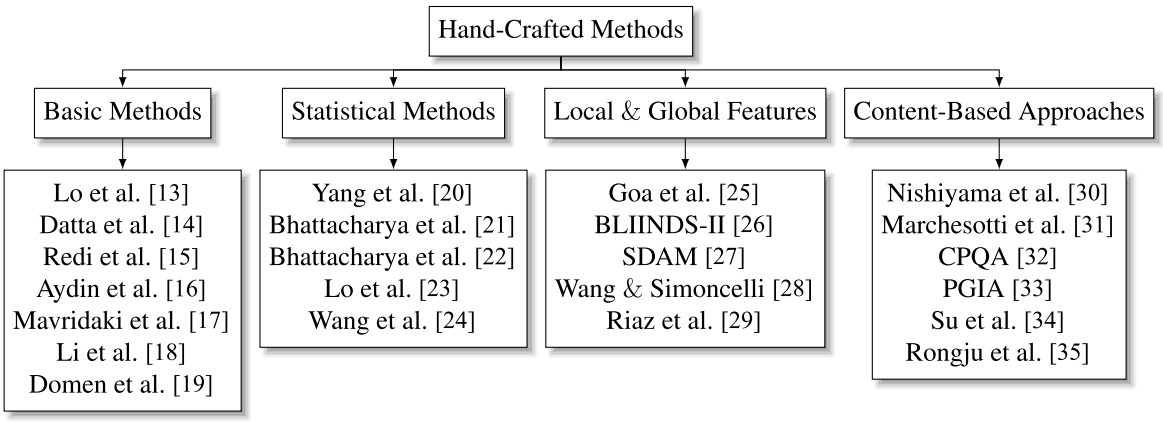



**FIGURE 2.** Overview and Categorization of hand-crafted based techniques for image aesthetics assessment.

score is predicted. First, the framework is tested on a small scale, and later it is tested on a large scale, classifying the images as beautiful and non-beautiful via SVM classification. The AVA dataset is used for training and testing, achieving 75.76% accuracy.

4) A photographic rating framework that computes aesthetic signatures using attributes of colorfulness, sharpness, depth, tone, and clarity as shown in Figure 3(b) is introduced by Aydin *et al.* [16]. Generally, relation to photographic rules and clear definition such as sharpness, clarity, colourfulness, etc. are essential building blocks of any image's aesthetic algorithm. In the framework, a picture is shown on a screen with five images displayed on other screens and a short task

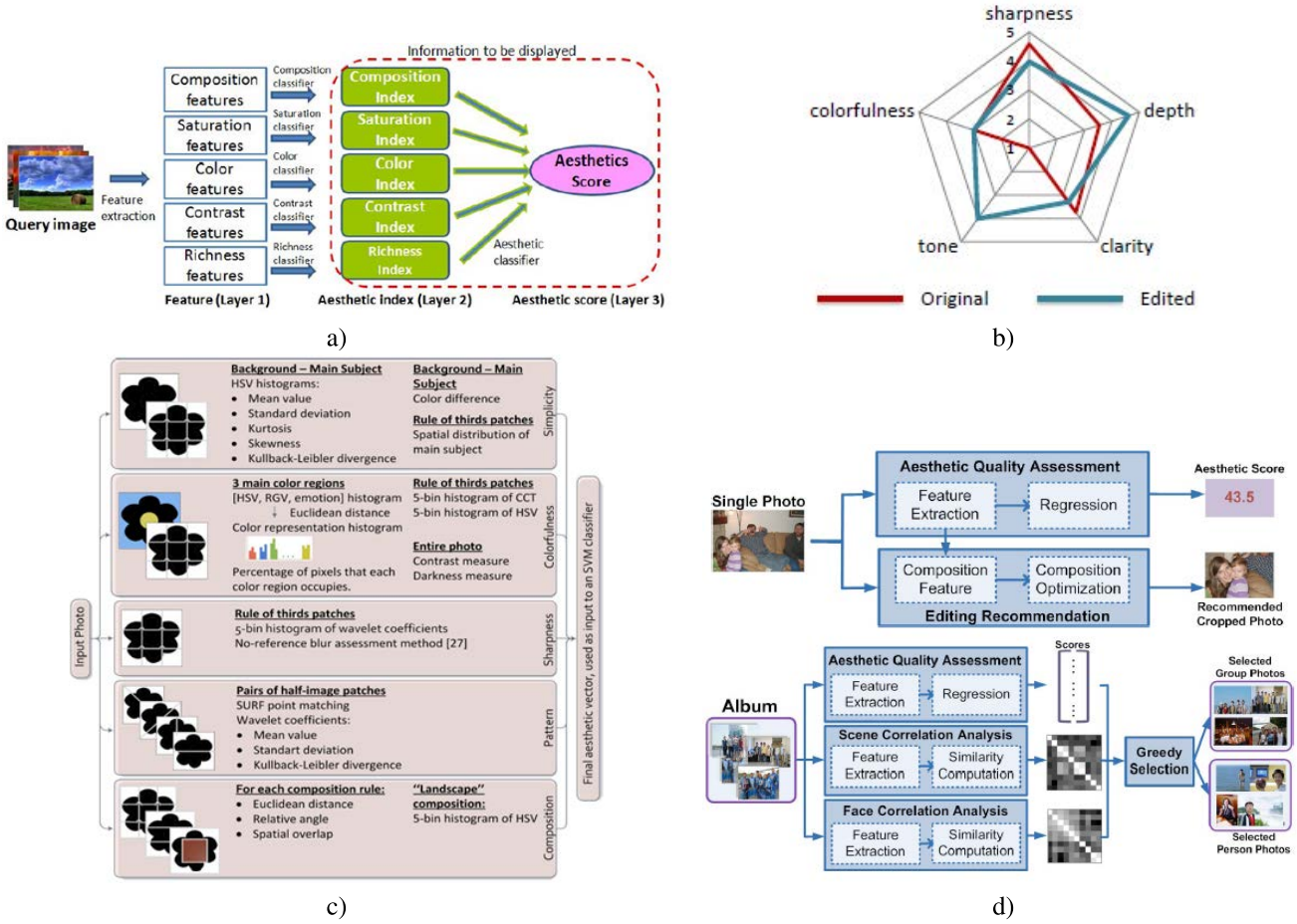



**FIGURE 3.** Comparative analysis of Basic hand-crafted assessment methods. a) Hierarchical Aesthetic Assessment Algorithm [13], b) Proposed Attributes in [16], c) Rule Based System [17], d) Photo Quality Assessment & Selection System [18].

description to determine the stimuli from the image for the five primary attributes, i.e., colourfulness, sharpness, depth, tone, and clarity where the task description contains an aesthetic rating of the image individually for each attribute, working on 8-bit RGB images. The algorithm works in three steps: i) convert the input image to a double-precision image and normalize it, ii) an edge pyramid is computed with domain transform applied to each pyramid layer, and iii) a multi-scale [42] contrast image is estimated. Moreover, a data structure is built using detailed contrast images, known as a focus map that indicates in-focus regions in the image, and the inverse of the focus map depicts out-of-focus image regions. The focus map is used to calculate features such as depth, colourfulness, sharpness, clarity, and tone. The training is performed on 955 images randomly selected from DpChallange [43] dataset. This research is applied mainly in HDR tone mapping, automatic photo editing applications, auto aesthetic analysis, and multi-scale contrast manipulation.

5) Mavridaki *et al.* [17] introduced a system using five basic photography rules: simplicity, colourfulness, sharpness, pattern, and composition. The *simplicity* refers to capturing images with emphasized subjects. For *colourfulness*, k-means clustering is performed to separate different colours. For *sharpness*, blur detection algorithm [44] is employed, and for *pattern* assessment, SURF point features [45] are extracted. For *composition* rule, landscape compostion [46], and rule of thirds [47] are examined. All these features are combined in the last stage to produce an element feature vector fed to an SVM classifier and are depicted in Figure 3(c). The mentioned method is evaluated on 12k images collected from CUHKPQ [48], CUHK [38], and AVA [49] datasets, where half of them are high-quality, and the other half are of low-quality images. The proposed framework achieves an overall accuracy of 77.08 %.
6) An online photo-quality assessment and photo selection system is present in [18] as shown in Figure 3(d), where the users post their images, and the algorithm provides aesthetic evaluation and editing recommendations. The cropping-based editing algorithm uses composition features and composition optimization

for the proposed system inputting a single image or photo album. The aesthetic score is calculated for a single image between 0-100, and image crop recommendations are provided if the aesthetic score is less than 70. Similarly, the top ten rated group photos and single-person photos are displayed in the photo album with respective scores. Aesthetic assessment is performed by extracting a feature vector from images followed by regression to compute the aesthetic score. Features are considered based on colour, light, composition, and face characteristics. The aesthetic quality assessment algorithm is trained on a dataset of 500 photos collected from Amazon Mechanical Turk with ten test images. For albums, images are categorized into single-person and group photos. Moreover, scene correlation analysis and face correlation analysis are performed for group photos, and aesthetic quality is determined.

7) Using feature extraction and SVM classifier, Domen *et al.* [19] proposes an aesthetic photo technique, where three basic photography features, including simplicity, composition, and colour selection, are considered for aesthetic assessment. The edge features determine simplicity and the ratio of background to image colour palette. The rule-of-thirds and golden ratio assess composition. To classify image in the high aesthetic score and low aesthetic score, the SVM classifier is trained on 258, and 1048 images are randomly selected from the Flicker and the DPChallange [43] datasets, respectively, achieving an accuracy of 95% using 73 features from each image.

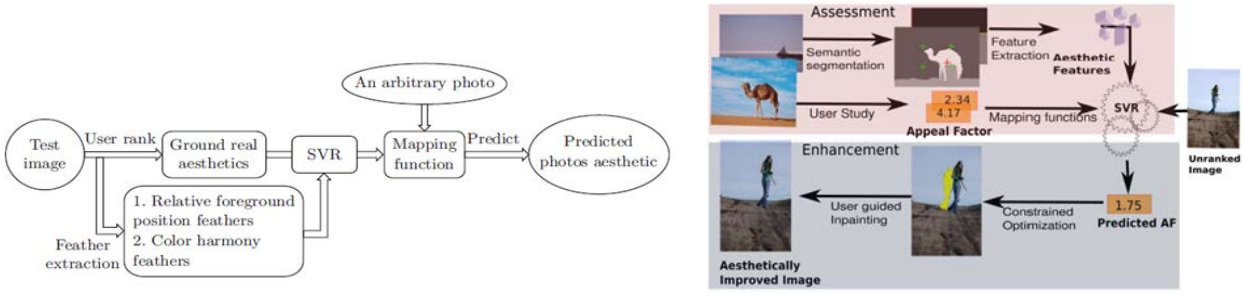
### B. STATISTICAL METHODS

In this subsection, we discuss various methods for image aesthetics based on the statistics of texture, foreground and background.

1) The landscape photo assessment algorithm by Yang *et al.* [20] is shown in Figure 4(a). The authors extract as relative foreground position and colour harmony features, and according to the rule of thirds, the object of interest must be at the image centre. Moreover, colour harmony is the relative position of each colour in the spatial domain, and colour harmonic normalization [50] is performed via hue wheel. The support vector regression (SVR) algorithm [51] is trained to map the foreground position and colour harmony features with the ground real aesthetics. A mapping model is learned to predict the aesthetic level after achieving the composition deviation and is tested on 431 images from Pconline and Flickr [52] concerning 84.83% accuracy.
2) Recently, a photo-quality assessment and enhancement algorithm to train SVR employing the relative foreground and visual weight ratio image features is given in [21], the architecture of their proposed framework is in Figure 4(b). The image is edited if the appeal factor is lower than the computed aesthetic score (i.e., between 1 to 5). The dataset consists of 384 single object images, and 248 images are scene images downloaded from Flickr [52]. This approach achieves 86% accuracy.
3) After image and photo aesthetic assessment [21], Bhattacharya *et al.* [22] next presented an aesthetic assessment framework for videos. As the algorithm deals with videos, three-level features are extracted, including cell level, frame level, and shot level. The *cell features* comprise dark-channel, sharpness, and eye sensitivity. For *frame-level*, Sentibank library [53] detects 1,200-dimensional feature vector. For *shot features*, the foreground motion [54], [55], [56], [57], the background motion and the texture dynamics [58] are computed from the video. A SVM is trained for each mentioned level feature. Finally, all SVM scores are fused using low-rank late fusion (LRLF) [59] while the algorithm is evaluated using NHK dataset [60] comprising of 1k videos.
4) Lo *et al.* [23] utilizes the colour palette, layout composition, edge composition, and global texture features for aesthetic assessment. The HSV histogram colour components extract colour palette features; layout composition features are determined through the $\ell_1$ distance between H, S, and V channels; edge detection filters compute edge composition features [61]; global texture features are calculated by the sum of absolute differences between four channels. In addition to the features mentioned above, blur, dark channel, contrast, and HSV counts are also computed. An SVM classifier trained on the CUHK dataset rates the image in high and low aesthetic levels, providing 86% accuracy in the performance.
5) Using saliency enhancement, Wang *et al.* [24] introduced an image aesthetic level prediction algorithm. The authors use the salient region of the image to represent objects, computing the saliency map via Ittiś visual saliency model [62]. The visual features from the image are extracted i.e., global, saliency regions, and foreground-background relationship features, where the global features are composed of texture details, low depth of fields, and rule of thirds. It is also to be noted that the distribution, position, and area of salient regions are determined as features of salient regions. The hue count and edge spatial distribution represent the foreground-background relationship. Moreover, images are classified into high and low aesthetic levels by an SVM classifier trained on a dataset downloaded from Photo.net [63], which contains 3161 images and achieves an accuracy of 83.7%.

### C. LOCAL AND GLOBAL FEATURES METHODS

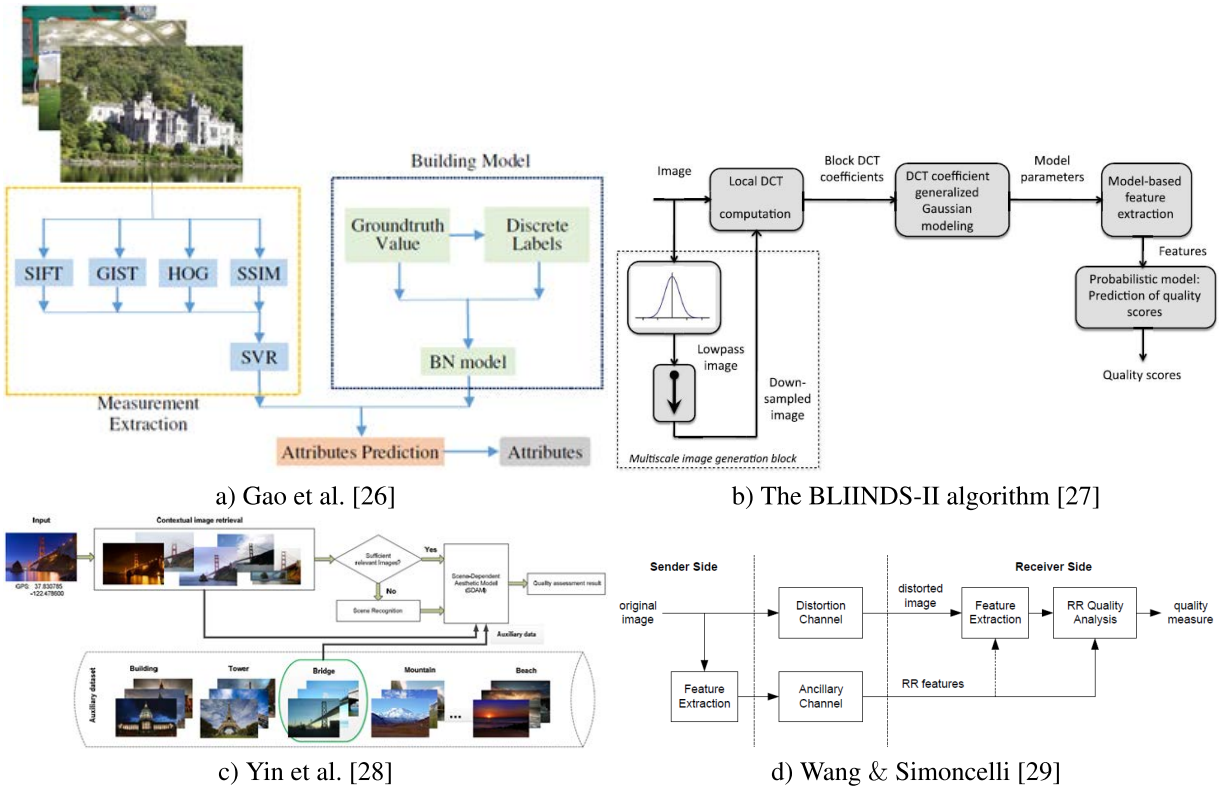
In this section, we summarize the algorithms that consider both local and global features learned from images for aesthetic assessment.




a) The Framework of Yang et al. [21]

b) The Method of Bhattacharya et al. [22]

**FIGURE 4.** Comparative analysis of hand-crafted statistical methods.




a) Gao et al. [26]

b) The BLIINDS-II algorithm [27]

c) Yin et al. [28]

d) Wang & Simoncelli [29]

**FIGURE 5.** The frameworks of local and global hand-crafted features algorithms.

1) The multi-label task for assessing the aesthetic quality of images based on different aesthetic attributes like aesthetic, memorable, and attractive attributes using high-level semantic information is explored in [25] as shown in Figure 5(a) by designing a Bayesian Network to predict the aesthetic level using multi aesthetic attribute prediction. Furthermore, a three-node Bayesian Network presents each aesthetic attribute, including its label, value, and measurement. There are two modules of the mentioned framework measurement acquisition by SIFT [64], GIST [65], HOG [66] or self-similarities and multi-attribute relation modeling. Finally, a support vector regression (SVR) is trained, the ground truth values are discretized in the building model, and a hybrid Bayesian Network structure is learned on continuous and discrete values. The training (with ten-fold cross-validation) and testing are performed on the memorability dataset [67] containing 2222 images and are evaluated on three different metrics: F1-score, Kappa, and accuracy.
2) The BLIINDS-II algorithm [26] employs discrete cosine transform (DCT) [68], [69] is given in Figure 5(b), where local DCT is computed utilizing input image and lowpass downsampled image. Afterwards, a gaussian model is built, extracting model-based features, which are then fed to a Bayesian model

that predicts the quality scores. The simple Bayesian probabilistic model requires minimum training [70] and is trained on randomly selected data samples from the LIVE IQA dataset [71] containing 779 images. The algorithm yields 91% accuracy.

3) A scene-dependent aesthetic model (SDAM) [27] takes into account both visual content and geo-context by utilizing the transfer learning [72] approach, where input images along with their geo-context (online images with similar contents as that of the input image) are used (see Figure 5(c)). The SDAM learns from two types of images, i.e., one category is geo contextual images that are location-wise similar to online photos, and in the other category, similar class images from the available database (DB). If a sufficient number of contextual images are available, the machine learning [73] approaches are applied to access the input image quality. The contextual image retrieval may contain the location of the same images but with different objects where the GIST identifies these types of irrelevant images and are discarded. Moreover, to learn, SDAM uses a state vector machine (SVM), which is tested on 9600 geo-tagged and 32k auxiliary dataset images, achieving an accuracy of 81% on popular spots and 73% accuracy on images of less prominent locations.
4) Wang and Simoncelli [28] uses a wavelet domain natural image statistical model, providing a distortion measure algorithm for communication systems where images are transferred from one location to other. The input image is decomposed into 12 wavelet bands, i.e., three scales and four orientations. The six wavelet bands are randomly selected to extract features and minimize KLD [74], rendering a quality score to rate images in different distortion levels. The architecture of the proposed deployment scheme is given in Figure 5(d). The framework is tested on a LIVE database containing 489 images showing 92% accuracy.
5) Recently, Riaz *et al.* [29] employs generic features, including both global and local features, by extracting SURF features in addition to wavelet and composition features. The method also determines basic photographic features, colour combination, saturation, contrast, smoothness, intensity, hue, and aspect ratios from the input image. The approach applied three steps; in the first step, the online database comprising 250 images ( downloaded from Photo.net), in the second step, human professionals rate pictures, and the third step, all the features mentioned above are extracted. An artificial neural network is trained on these features, achieving 83% accuracy.

### D. CONTENT-BASED METHODS

Content-based methods take into account the content of the images. We provide an overview of such methods in the following paragraphs.

1) Aesthetic quality is highly based on the local region's sum of colour harmony scores according to Nishiyama *et al.* [30], implementing bag-of-colour patterns for photograph quality classification (see Figure 6(a)). The authors employ the moon and spencer model [75], [76] computes the sum of colour harmony scores. The colour model evaluates the hue, chroma, and lightness from the sampled local regions of images, and then the collected distributions are integrated to form a bag-of-features [77] framework. Every local area is described using simple colour patterns of colour harmony models, assuming the colour distribution to be simple. Aesthetic rating is classified by calculating the histogram of each colour pattern. The SVM classifier is trained on 124,664 images collected from the DPChallange [43] dataset to predict the photograph quality, categorized into high and low aesthetic levels. The algorithm is tested in two scenarios: a) whole image and 2304 local regions each of size $32 \times 32$, and (b) absolute and relative colour values offer an overall 77.6% accuracy. To further improve the classification, the authors also consider the saliency, blur, and edge features in addition to colour harmony patterns.
2) Marchesotti *et al.* [31] proposed an image descriptor-based with Fisher vector (FV) [78] and bag-of-visual-words (BOV) [79] which extracts generic descriptors from image and gradient information is obtained through Scale Invariant Feature Transform (SIFT). The input image is divided into patches, and for each patch, BOV computes discrete distribution, and FV calculates continuous distribution. Furthermore, SIFT is applied to each patch, and GIST descriptor is also considered, initially designed for scene categorization. The algorithm is evaluated on two datasets, Photo.net and CHUK, consisting of 3581 images and 12k images, respectively. The BOV and FV features are computed from $32 \times 32$ patches at five different scales and represented by SIFT that generates a 128-dimensional feature vector for each patch reduced to 64 dimensions using PCA. The EM [80] algorithm learns visual vocabulary Gaussian mixture models, and the SVM classifier is learned using hinge loss and stochastic gradient descent algorithm [81], [82]. In their experiments, Fisher Vector outperforms all other techniques and delivers a maximum of 78% accuracy.
3) The content-based photo quality assessment, abbreviated as (CPQA) [32], deals with both regional and global features concerning three different areas, including clarity-based detection, layout-based detection, and human-based detection. Regional features extracted from the input image are dark channel, clarity-contrast, lighting-contrast, composition geometry, complexity, and brightness. Besides, the global features include hue and scene composition features. An SVM is trained on the CUHK-PQ [48], including 17673 images

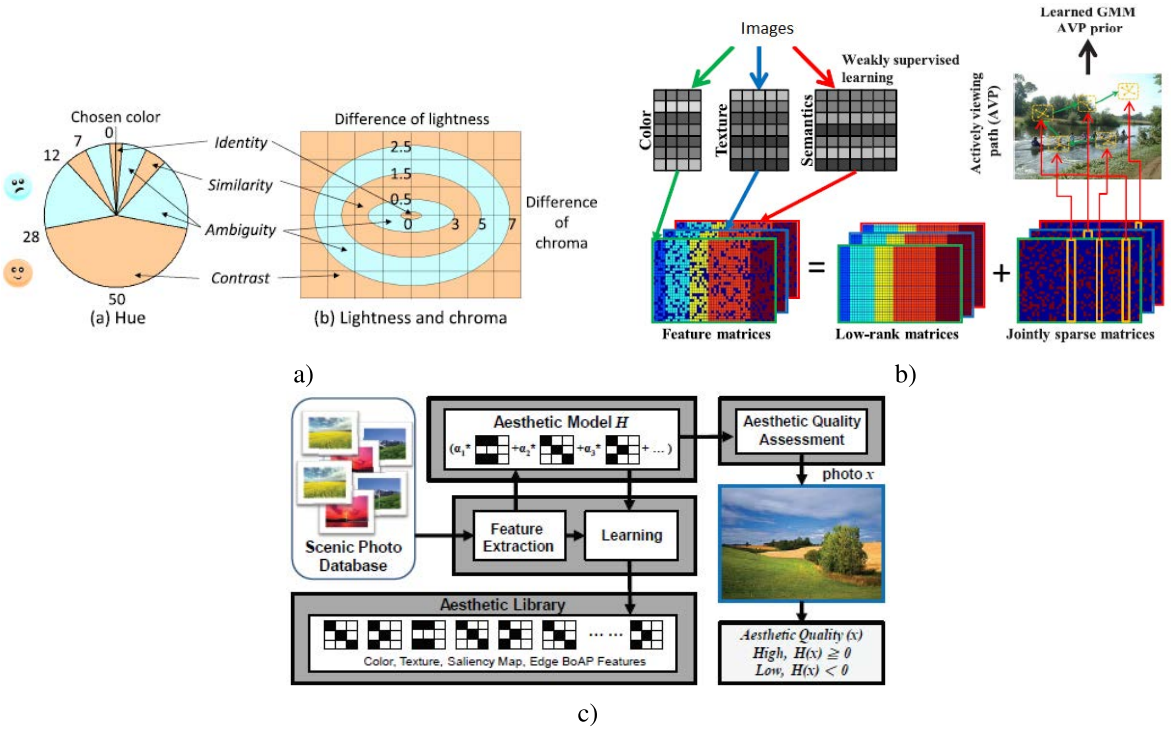



**FIGURE 6.** **Comparative analysis of content-based Methods. a) Nishiyama *et al.* [30]'s Moon and Spencer Model, b) Algorithm by Zhang *et al.* [33], and c) Su *et al.* [34]'s algorithm overview.**

classifying the images into high, low, and uncertain categories. The CPQA algorithm gives 83% accuracy.

4) The perception-guided image aesthetic (PGIA) [33] assessment algorithm learns the model constrained with different low-rank graphlets created by fusing low-level and high-level features from the image. The sparsity of the graphlets is then calculated to generate jointly sparse matrices as shown in Figure 6(b). The mentioned graphlets turn into actively viewing path (AVP) descriptors, and the Gaussian Mixture Model learns the distribution of these aesthetic descriptors. The proposed algorithm is trained and tested on AVA [49], Photo.net, and CUHK datasets comprising of 12k, 3581 images, and 25k images, providing 90.59%, 85.52%, and 84.13% accuracy, respectively.
5) Su *et al.* [34] proposes a bag-of-aesthetics preserving (BoAP) library. The algorithm is implemented in two steps: 1) The image is decomposed into multiple resolutions, 2) extraction of bag-of-aesthetics features. The HSV colour space, local binary patterns, and saliency map extract features from the images. The AdaBoost classifier is trained and tested on a dataset of 3k images downloaded from DPChallange [43] and Flickr [52], providing 92.06% accuracy. Figure 6(c) shows the framework of the BoAP algorithm.
6) To evaluate the quality of Chinese handwriting, Rongju *et al.* [35] explored the problem of artificial intelligence with aesthetic feature representation. The first step extracts component layout features and global features from Chinese handwritten images. For components, the semi-automatic component extraction method extracts layout feature strokes. Similarly, the alignment, stability, and distribution of white spaces and gaps between strokes are global features extracted from input Chinese handwritten images. A novel dataset named Chinese Handwriting Aesthetic Evaluation Database (CHAED) [35] is built and used to train the SVM classifier. Finally, neural networks are trained on CHAED to get the aesthetic evaluation ability.

## III. DEEP LEARNING METHODS

Deep learning uses artificial neural networks to automatically learn complex low and high-level features useful for computer vision tasks [95], [96]. In many cases, deep learning has produced results comparable to human accuracy or even surpassed humans in many areas. Convolutional neural networks are the backbone of deep learning for image analysis [97], [98], [99]. Once trained on millions of images, these networks can provide outstanding accuracy on image understanding tasks such as image aesthetic assessment. In this section, we discuss various important works for image aesthetics prediction using deep learning methods, shown in Figure 7.

The aesthetic quality assessment of photographs can be formulated as a classification or regression or the combination of classification and regression approaches. There is a lack of consensus on the definition of aesthetic quality as it is a subjective matter. However, the photo-sharing communities rated the photos, and the average score is usually taken as the

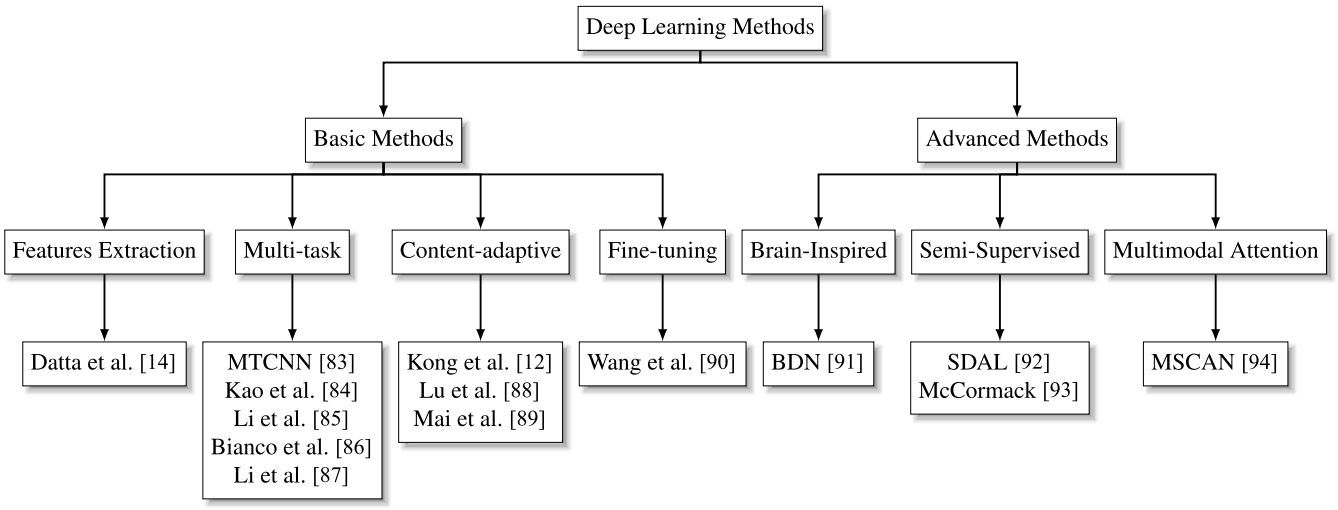



**FIGURE 7.** Overview of the deep learning methods and their classification based on similarity in structure.

quality of the images and used as ground truth for different algorithms. Therefore, the quality of the assessment task is taken as a classification problem. Nevertheless, the problem can also be formulated as a regression to regress the quality of photographs to aesthetic score. Thus, the aesthetic quality assessment feature could be either extracted as a hand-crafted or learned using deep learning architecture in multi-task settings. The multi-task approaches tend to learn better and improve the aesthetic score significantly.

### A. DEEP LEARNING BASIC METHODS

#### 1) DEEP FEATURES EXTRACTION BASED METHODS

The author used a deep neural network and extracted 56 visual features originally proposed by Datta *et al.* [14] for aesthetic assessment [100]. The dataset is collected from the internet consisting of 28896 images, where each image is resized to $160 \times 120$ resolution, and features are extracted from images by converting them to HSV colour space. These extracted features include brightness, Earth Mover Distance (EMD), Hue, saturation, etc. The autoencoder has been used to compress the raw features into new features with $1/2$ or $1/4$ of its original input. The Artificial Neural Network (ANN) trains the network with both the extracted 56 visual features and the new features extracted from the autoencoder. The ANN is used here as a classifier to classify the photograph into two aesthetic categories: high and low aesthetics. Moreover, Convolutional Neural Networks and Deep Belief networks are also used for aesthetic evaluation on a larger dataset. The overall structure of their proposed scheme is depicted in Figure 8. This scheme is tested on an AMD Athlon II PC providing 82.1% accuracy. A global, local, and scene-aware information of images are considered and exploited the composite features extracted from corresponding pre-trained deep learning models for classification using SVM [101]. They found that a deep residual network could produce more aesthetics-aware image representation and composite features.

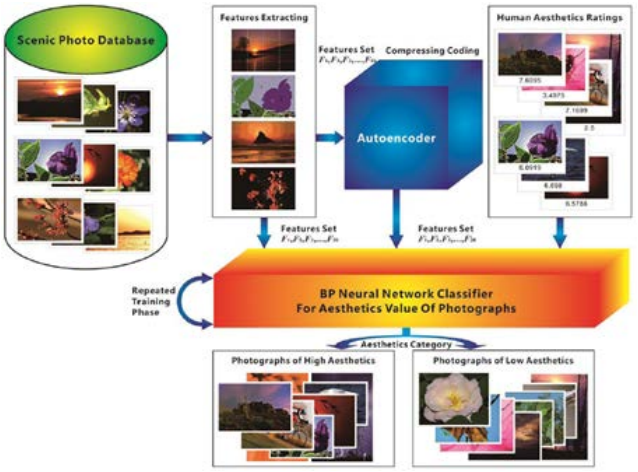



**FIGURE 8.** The scheme for deep features extraction methods [100].

#### 2) MULTI-TASK CONVOLUTIONAL NETWORKS

A Multi-task learning approach is employed to explore the correlation between automatic aesthetic assessment and semantic information. The task is to utilize semantic information in the joint objective function to improve the quality assessment task [102]. The approach provides both aesthetic and semantic labels as output. A Multi-Task Convolutional Neural network (MTCNN) [83] is designed that performs both semantic recognition and quality assessment considering an input image size of $227 \times 227$. The proposed CNN automatically learns the relation between semantics and aesthetics. Their CNN consists of five convolutional layers, three pooling layers, and three fully-connected layers. The proposed Convolutional Neural Network architecture is shown in Figure 9(a). Furthermore, three representations

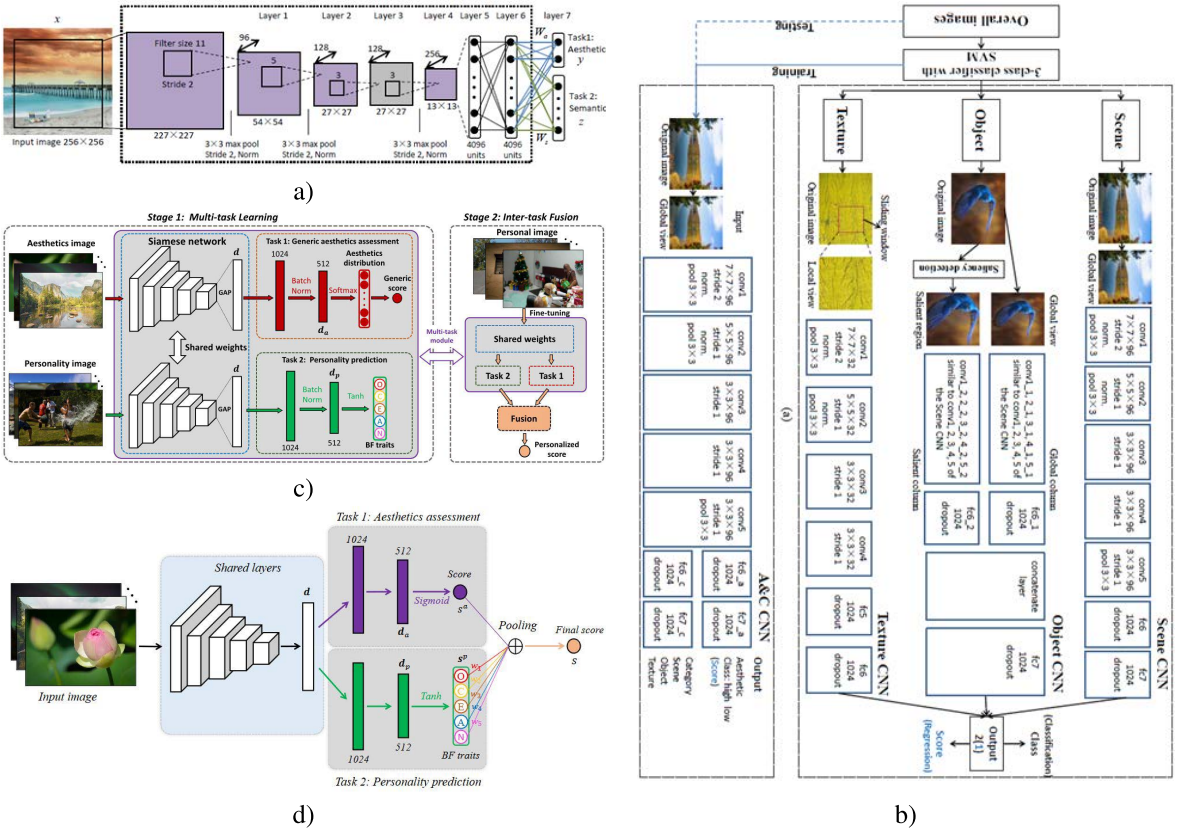



**FIGURE 9.** Multi-task Convolutional Networks Based Methods. a) Architecture of the system proposed by Kao *et al.* [102], b) Framework of the system proposed by Kao *et al.* [84], c) The personality-assisted multi-task learning model by Li *et al.* [87], and d) The architecture of the multi-task learning model by Li *et al.* [85].

of the MTCNN are proposed in which different configurations of convolutional layers and pooling layers [103], [104], [105] are designed. A multi-task probabilistic framework is applied. The network is trained and tested on the AVA dataset [49] and Photo.net [14] dataset. AVA dataset consists of 255k images, and the photo.net dataset comprises 20,278 images. On the AVA dataset, MTCNN achieves up to 77.71% accuracy, and on Photo.net, it achieves up to 65.20% accuracy.

A Convolutional Neural Network-based framework has been proposed for the visual quality assessment [84]. There are three categories defined for each image; scene, object, and texture. Firstly, each image is classified into one of the three categories using SVM. Then for each category, a separate convolutional neural network named Scene CNN, Object CNN, and Texture CNN is trained to learn features and classify the output into a high aesthetic or low aesthetic class and a numerical aesthetic score. In addition, another single CNN called A&C CNN is deployed, which performs recognition of quality and aesthetic ratings simultaneously for overall images. Figure 9(b) shows the overall structure of the implemented scheme. The algorithm is tested on an AVA dataset containing 255k images. It achieves 91.3% accuracy. The scene, object, and texture CNN are highly dependent on the classification accuracy of the SVM classifier. If SVM provides the wrong classification, the incorrect CNN gets activated and outputs inaccurate results.

An end-to-end personality-driven multi-task deep learning model has been introduced to assess the aesthetics of an image [85] as shown in Figure 9(c). Firstly, image aesthetics and personality traits are learned from the multi-task model. Then the personality features are used to modulate the aesthetics features, producing the optimal generic image aesthetics scores.

Bianco *et al.* [86] used deep learning to predict image aesthetics using aesthetic visual analysis (AVA) [49] dataset. This model fine-tuned canonical convolutional neural network architecture to obtain aesthetic scores in this model. Aesthetic quality assessment is treated as a regression problem. Caffe network [106] is selected to be fine-tuned, and the last fully connected layer of CaffeNet is replaced by a single neuron providing an aesthetic score between 1 and 10. Another modification is incorporated in Caffe Net to use Euclidean loss [107] instead of Softmax loss [108]. A stochastic gradient descent backpropagation algorithm fine-tunes the new network. The dataset contains 255k images, from which 250,129 images are used for training and 4970 images for testing. The algorithm achieves 83% accuracy.

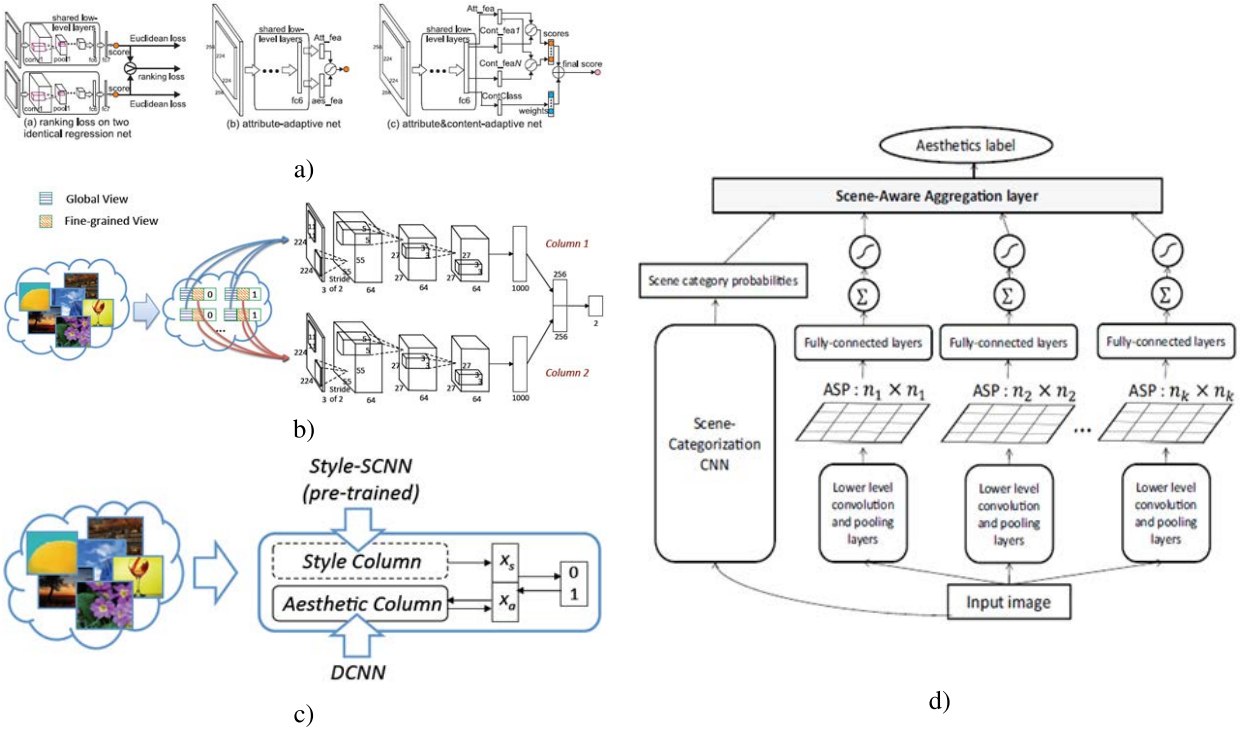



**FIGURE 10.** Content-adaptive Deep Learning Methods. a) Architectures of different models proposed by Kong *et al.* [12], b) Double column convolutional neural network model for aesthetic quality assessment proposed by Lu *et al.* [88], c) Regularized double-column convolutional neural network model proposed by Lu *et al.* [88], and d) Scene aware multi-net regression model of Mai *et al.* [89].

A personality-assisted multi-task deep learning framework is presented [87] as shown in Figure 9(d) for both generic and personalized image aesthetics assessment. Initially, they introduced a multi-task learning network with shared weights to predict the aesthetics distribution of an image and Big-Five (BF) personality traits of people who like the image. They then used an inter-task fusion to generate individuals' personalized aesthetic scores on the image.

### 3) CONTENT-ADAPTIVE DEEP LEARNING METHODS

A content adaptation technique using deep CNN has been proposed for image quality aesthetic assessment [12]. A new dataset is published by these researchers, which they named as Aesthetics and Attributes Database (AADB) [12] comprising 10k images. AlexNet architecture [109] is fine-tuned on AADB dataset. Softmax loss is replaced by Euclidean loss. Another Siamese network [110], [111] is fine-tuned with content category classification and attribute layers to achieve hybrid performance. An attribute-adaptive model and a content-adaptive model are designed. Figure 10(a) shows three different models initially based on AlexNet. Model (a) uses shared low-level layers of AlexNet and adopts Euclidean loss and Ranking loss, whereas model (b) is an attribute-adaptive net with an additional attribute predictor branch. Model (c) provides a combined adaptive net and attribute adaptive net approach. It takes an input image of size $227 \times 227$ and provides 77.33% accuracy.

A two-column content-adaptive aesthetic rating neural network is proposed that takes into account both style contents and semantic information [88]. Each column is trained on two different crops of a single image. Each column consists of three convolutional layers and three pooling layers followed by a fully connected layer. Finally, style and semantic features extracted by both columns are fused by two fully connected layers as shown in Figure 10(b). The network is trained using end-to-end learning and stochastic gradient descent. A network adaptation strategy is proposed to facilitate content-based image aesthetics. This helps improve the adaptation of images' semantic contents; hence, fewer images from each category are required for training. A Regularized Double-column Convolutional Neural Network (RDCNN) is proposed, which includes a single Style Column Convolutional Neural Network (Style-SCNN) for style information and a Double-Column Convolutional Neural Network (DCNN) for semantic information. The final structure of the framework is shown in Figure 10(c). This network is tested on the AVA dataset and IAD dataset [112] to categorize images into high and low quality and achieves 71.2% accuracy.

A composition preserving convolutional neural network has been proposed for photo aesthetic assessment [89]. The network incorporates the concept of image quality degradation by resizing and clipping. Multi Net Adaptive spatial pooling Convolutional Neural Network (MNA-CNN) is designed to rate variable size images. For this purpose, an adaptive spatial pooling layer is introduced that adjusts its receptive size according to output rather than input. There are multiple streams of network [113] where an adaptive spatial pooling layer replaces the last pooling layer. Pre-trained VGG [114] is

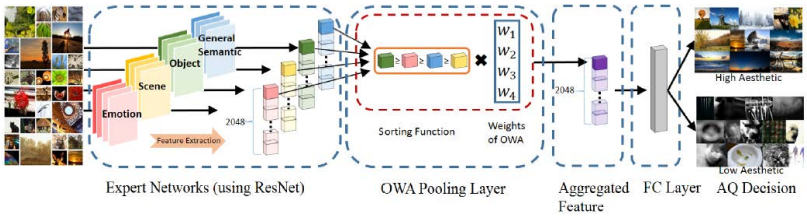



**FIGURE 11.** Deep semantic aggregation network proposed by Lu *et al.* [116].

fine-tuned on Torch Deep Learning package [115], and each sub-network is trained separately. Another scene categorization CNN is trained on Places205-GoogleLeNet consisting of 2.5 million images. This framework is shown in Figure 10(d). Scene categorization network increases aesthetic assessment accuracy to 77.1% accuracy.

#### 4) FINE-TUNING BASED APPROACHES

A pre-trained convolutional neural network is fine-tuned for assessing the quality of images [90]. AlexNet and VGG are fine-tuned to provide output in two categories (high and low). VGG is a deeper network than AlexNet, providing high accuracy and requiring more training time. AlexNet comprises five convolutional layers with ReLU non-linearity, five pooling layers, and three fully-connected layers. The last layer is replaced by a fully connected layer for a two-class classification. VGG is a deeper network consisting of sixteen to nineteen convolutional and pooling layers. Both global and local views train the networks. AVA and CUHKPQ datasets are used to fine-tune and are trained on both the global and local views. AlexNet achieves 91.20% accuracy CUHKPQ dataset, and VGG achieves 91.93% accuracy. AlexNet achieves 83.24% accuracy on the AVA dataset, and VGG achieves 85.41% accuracy.

A ResNet152 network has been used for image aesthetic quality assessment [116], which was trained on the ImageNet dataset for object classification and further fine-tuned on AVA, Places, and emotion6 datasets. The network is trained for four different categories; scene images, object images, emotion images, and general semantic images as depicted in Figure 11. For the scene images, 2.5 million images from the Places dataset [117] are used to fine-tune ResNet152. The network is trained using the AVA dataset for object images, and the emotion images network is trained on the Emotion6 dataset consisting of 1980 images. This network achieves 78.6% accuracy.

### *B. ADVANCED DEEP METHODS*

#### 1) BRAIN-INSPIRED APPROACHES

A Brain-inspired Deep Neural Network (BDN) has been proposed for image aesthetic assessment [91] and is composed of two parts. The first part is attribute learning via parallel pathways, and the second part is a high-level synthesis network as shown in Figure 12(a). Attribute learning via parallel pathways is a combination of deep neural network streams. Different attributes are learned from input images, including hue, saturation, value, complementary colours, duotones, high dynamic range, image grain, light on white, long exposure, macro, motion blur, negative image, rule of thirds, shallow DOF, silhouettes, soft focus and vanishing point. Hue, saturation, and value are directly computed from the image, whereas the other attributes are learned using parallel deep neural networks as shown in Figure 12(b). This network predicts a label 0 or 1 and is trained using the AVA dataset. Their high-level synthesis network is a four-layer convolutional neural network. This network predicts the overall aesthetic level of the image. At this stage, the entire network is trained end-to-end using the AVA dataset. Experiments are performed on 12 CPUs (Intel Xeon 2.7 GHz) and a GPU (Nvidia GTX680). Training and fine-tuning take around one day with an accuracy of 76.80%.

#### 2) SEMI-SUPERVISED APPROACHES

For image aesthetic quality assessment, Liu *et al.* [92] proposed a semi-supervised deep active learning (SDAL) algorithm, which discovers how humans perceive semantically significant regions from many images partially assigned with contaminated tags.

An adaptive fractional dilated convolution is developed [118], which is aspect-ratio-embedded, composition-preserving and parameter-free. The fractional dilated kernel is adaptively constructed according to the image aspect ratios, where the interpolation of the nearest two integers dilated kernels are used to cope with the misalignment of fractional sampling.

A convolutional neural network is used to investigate the relationship between image measures, such as complexity, and human aesthetic evaluation, using dimension reduction methods to visualize both genotype and phenotype space to support the exploration of new territory in a generative system [93]. Convolutional neural networks trained on the artist's prior aesthetic evaluations are used to suggest new possibilities similar to or between known high-quality genotype-phenotype mappings.

#### 3) MULTIMODAL ATTENTION-BASED NETWORKS

The MSCAN, a multimodal self, and collaborative attention network is proposed for aesthetic prediction task [94] as shown in Figure 13. The self-attention module finds the response at a position by attending to all positions in the

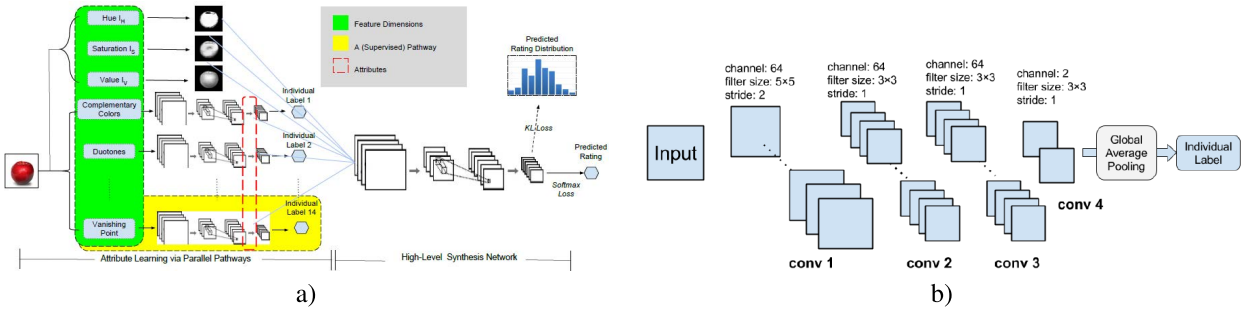



**FIGURE 12.** Brain-inspired Approaches of [91]. a) Brain inspired network architecture and b) Attribute learning via parallel pathways.

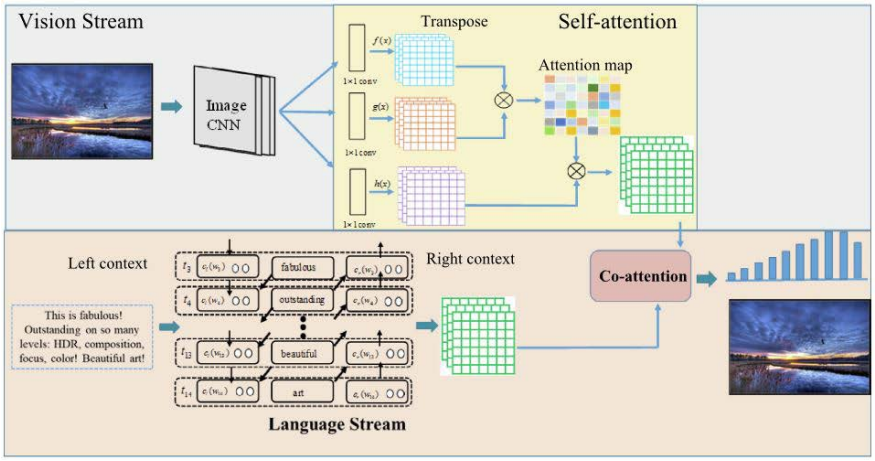



**FIGURE 13.** The multimodal self and collaborative attention network by Zhang *et al.* [94].

images to encode spatial interaction of the visual elements. To model the complex image-textual feature relations, a co-attention module is used to perform textual-guided visual attention and visual-guided textual attention jointly.

## IV. EXPERIMENTAL SETTINGS

To make the survey more comprehensive, we first provide information about the publicly available widely used benchmark datasets and evaluation metrics, followed by the hand-crafted and deep learning comparisons.

### A. DATASETS

#### 1) PHOTO.NET

This dataset [14] is collected from ''Photo.net'', a website of photo-sharing community established in 1997. The authors considered originality and aesthetic qualities used for rating photos on this website. Both the qualities are correlated, but originality is considered by the authors to be used for further processing due to its role in aesthetic value. The authors finally obtained 3581 photos for their work. The original dataset contains 20278 images.

#### 2) AESTHETIC VISUAL ANALYSIS DATASET

Aesthetic Visual Analysis (AVA) dataset [49] is derived from ''dpchallenge.com'' where the community uploads images to participate in different photographic challenges having titles and descriptions. In this connection, each image is linked with the information of its corresponding challenge that can provide the context of annotations when combined with aesthetic scores or semantic labels.

AVA dataset contains 255,000 images that are associated with 963 challenges. While treating the aesthetic quality as a binary-class classification problem, images having an average aesthetic score value greater than the threshold value $5 + \sigma$ are labelled as positive. In contrast, those with an average aesthetic score value less than $5 - \sigma$ are negative. Training and testing sets contain 230,000 and 20,000 images respectively for a hard threshold $\sigma = 0$. Another split is also used to account for the top 10% and bottom 10% of the images, thus obtaining 25,000 images in the training set and 25,000 in the testing set.

#### 3) CUHK

CUHK [31] is a publicly available dataset that contains photos of diversified aesthetic quality where 60,000 images were collected from ''dpchallenge.com'' each of which is rated by a minimum of 100 users. The images with top 10% average rates are considered good category whereas the bottom 10% average rates are considered bad category and, therefore, are manually examined. Due to the fact that CUHK draws a clear

boundary between the classes, it is not a challenging dataset compared to the datasets where the class boundaries are not very clear.

#### 4) CUHK-PHOTOQUALITY

CUHK-PhotoQuality (CUHK-PQ) dataset [48] is a collection of 17,690 images obtained from multiple online community platforms and university students. The images are aesthetically labelled either as high quality or low quality based on the feedback of independent viewers. The label for each image is decided only if eight reviewers out of ten favours it. CUHK-PQ dataset covers seven distinct categories: animal, plant, night, human, landscape, architecture, and static. The data is randomly partitioned according to 50-50 split to generate training and testing sets where the ratio of positive to negative samples is 1:3.

#### 5) MIRFLICKR

In the domain of multimedia retrieval, MIRFLICKR dataset [49] is a collection of one million images accompanied by textual tags, aesthetic annotations in the form of Flickr's interestingness, and EXIF metadata. As opposed to the AVA dataset, the MIRFLICKR dataset has an interestingness flag only that describes the aesthetic preference. Exposure and blur are two aspects associated with 44 visual concepts in the MIRFLICKR dataset. Images in this dataset are categorized in the following categories: neutral illumination, over-exposed, under-exposed, motion blur, no blur, out of focus, and partially blurred.

#### 6) AESTHETICS AND ATTRIBUTES DATABASE

Aesthetics and Attributes Database (AADB) [12] is constructed by downloading 10k images from the Flickr website, where each image is rated by five raters independently. In this way, each image in the dataset is annotated with an aesthetic score and eleven attributes. The training, validation, and testing sets contain (8,500), (500), and (1,000) images, respectively. This dataset is distributed in different categories by the K-means clustering technique, where the value of k is set to ten based on experimental observation.

### B. EVALUATION METRICS

The most commonly used evaluation metrics in image aesthetic assessment are summarized in subsections.

#### 1) OVERALL ACCURACY

Overall accuracy (OA) takes into account True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) samples of a dataset. Accuracy may be misleading in the case of imbalanced data. However, it is a widely used measure to assess the performance of a classification model. It can be expressed mathematically as

$$OA = \frac{(TP + TN)}{(TP + FN + TN + FP)} \times 100 \tag{1}$$

#### 2) BALANCED ACCURACY

In the case of an imbalanced dataset, *Balance Accuracy* (BA) can be used to evaluate the performance of a classifier and averaging recall values can calculate it for each class. Balance accuracy is computed as the arithmetic mean of sensitivity and specificity. Mathematically, it can be expressed by Eq. (2).

$$BA = \frac{Sensitivity + Specificity}{2} \tag{2}$$

*Sensitivity* is the true positive rate that computes the correctly predicted positive samples out of total positive samples, whereas specificity is the true negative rate that computes the correctly predicted negative samples out of total negative samples. Sensitivity and specificity are given below

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

#### 3) PRECISION-RECALL CURVE

When the classes are highly imbalanced, the precision-recall curve is beneficial in assessing the performance of a classification model. The precision-recall curve highlights the trade-off between precision and recall for various threshold values. Higher the value of area under the precision-recall curve, higher are the values of recall and precision where the high precision value indicates a low false-positive rate and high recall value shows a low false-negative rate

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{(TP + FN}, \tag{6}$$

where precision and recall are given in Eq. (5) and (6), respectively. Here, precision refers to the number of true positives over the number of true positives and the number of false positives predicted by the classifier. On the other hand, recall indicates the number of true positives over the total number of positives, including true positives and false negatives in the positive class.

### C. ANALYSIS

This section provides a comparative performance analysis of both hand-crafted and deep learning-based methods. We provide an overview of how the reviewed techniques are different from each other with respect to features utilized, accuracy, dataset size, and classifiers used.

#### 1) PERFORMANCE OF HAND-CRAFTED METHODS

In this section, we show the analysis of various hand-crafted techniques as follows

- **Basic Feature Methods.** Table 1 presents the accuracy of basic feature methods results for image aesthetic assessment. Depending on the dataset used for testing,

**TABLE 1. Comparative analysis of hand-crafted techniques using basic image features for image aesthetic assessment.**

| Authors | Year | Features | | | | | | | | | | | | Classifiers | | | Dataset | | | | | | | No. of Images | Categorization Task | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Color | Hue | Saturation | Composition | Richness | Contrast | Brightness | Simplicity | Sharpness | Texture | Depth & Clarity | Tone | SVM | Linear Regression | Sparse Coding | PhotoWeb | DPChallenge | Flicker | Photo Database | CUHKPQ | CUHK | AVA | Self Collected | | level | |
| Ditta et al. [14] | 2006 | ✓ | ✓ | ✓ | | | | | | | | | | ✓ | | | ✓ | | | | | | | | 3581 | Bi | 71% |
| Li et al. [18] | 2010 | ✓ | | | | | ✓ | ✓ | | | | | | | ✓ | | | | | | | | | ✓ | 510 | Multi | 72% |
| Gadde et al. [119] | 2011 | ✓ | | | | | ✓ | ✓ | | | | | | ✓ | | | | ✓ | | | | | | | 12k | Bi | 79% |
| Pogacnik et al. [19] | 2012 | ✓ | | | ✓ | | | | ✓ | | | | | ✓ | | | | ✓ | ✓ | | | | | | 1306 | Bi | 95% |
| Lo et al. [13] | 2013 | ✓ | | ✓ | ✓ | ✓ | | | | | | | | ✓ | | | | | | ✓ | | | | | 9651 | Bi | 89% |
| Mavridaki et al. [17] | 2015 | ✓ | | | ✓ | | | | | ✓ | | | | ✓ | | | | | | | ✓ | ✓ | ✓ | | 12k | Bi | 77.1% |
| Redi et al. [15] | 2015 | ✓ | | | | | ✓ | | | ✓ | ✓ | | | ✓ | | ✓ | | | | | | | ✓ | | 250k | Bi | 75.7% |
| Aydin et al. [16] | 2015 | ✓ | | | | | | | | ✓ | | ✓ | ✓ | ✓ | | | | ✓ | | | | | | | 955 | Bi | - |

**TABLE 2. Comparative analysis of hand-crafted techniques using texture and FG/BG features for image aesthetic assessment.**

| Methods | Year | Features | | | | | | | | | | Classifiers | | Dataset | | | | | Dataset Size | Task | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Color | Relative FG Position | Visual Weight Ration | Sentibank | FG/ BG | Spatio-Temporal Binary Patterns | Global Texture/features | Edge Composition | Layout Composition & Color Palette | Salient Regions | SVR | SVM | Internet | Flickr | NHK Video Database | CUHK | Photo.net | | Level | |
| Yang et al. [20] | 2015 | ✓ | ✓ | ✓ | | | | | | | | ✓ | | ✓ | | | | | 431 | Multi | 84.8% |
| Bhattacharya et al. [21] | 2010 | | ✓ | | | | | | | | | ✓ | | | ✓ | | | | 632 | Bi | 87.3% |
| Bhattacharya et al. [22] | 2013 | | | | ✓ | ✓ | ✓ | | | | | | ✓ | | | ✓ | | | 1k | Bi | - |
| Lo et al. [23] | 2012 | ✓ | | | | | | ✓ | ✓ | ✓ | | | ✓ | | | | ✓ | | | Bi | 86.0% |
| Wang et al. [24] | 2010 | | | | | ✓ | | ✓ | | | ✓ | | ✓ | | | | | ✓ | 3161 | Bi | 83.7% |

**TABLE 3. Comparative analysis of hand-crafted techniques using local and global features for Bi-level image aesthetic assessment categorization task.**

| Authors | Year | Features | | | | | | Classifiers | | | | | | | Dataset | No. of Images | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | SIFT | GIST | SDAM | DCT | Wavelet | SURF | BN | SVR | KNN | Bayesian | SVM | KLD | ANN | | | |
| Gao et al. [25] | 2015 | ✓ | ✓ | | | | | ✓ | ✓ | | | | | | Proprietary | 2222 | 72.7% |
| Yin et al. [27] | 2012 | | ✓ | ✓ | | | | | | ✓ | | ✓ | | | Flicker | 10200 | 73% |
| Saad et al. [26] | 2012 | | | | ✓ | | | | | | ✓ | | | | Live | 799 | 91% |
| Wang et al. [28] | 2005 | | | | | ✓ | | | | | | | ✓ | | Live | 489 | 92% |
| Riaz et al. [29] | 2012 | | | | | ✓ | ✓ | | | | | | | ✓ | Photo.net | 250 | 83% |

the accuracy varies significantly. The maximum accuracy is obtained by [19] for DPChallenge and Flicker datasets. However, recent methods such as [15] and [17] reports less accuracy as the datasets employed are different and the number of images is significantly higher. The basic feature method's popular choice for the classifier is SVM, and the essential feature is colour.

- **Statistical Methods.** A comparison of accuracy, dataset size, classifier, and features extracted for statistical methods is given in Table 2. In multi-level classification, Yang *et al.* [20] achieves the highest accuracy of 84.83% while in bi-level classification Lo *et al.* [23] obtained 86%. Although the authors employed differents for each classification level. SVM is mostly employed for classification.
- **Global and Local Features Methods.** The local and global features methods are provided in Table 3 showing the comparison of accuracy, datasets, the number of images, classifiers, year of publication, and attributes extracted. The accuracy for the mentioned methods ranges from 72.7% to 92%. The algorithms utilize different features, datasets, and classifiers for each technique.
- **Content-Based Methods.** Table 4 gives the comparison between content-based hand-crafted methods. All the methods are evaluated for bi-level image aesthetic assessment tasks. The number of images employed by content-based is relatively higher than other previously mentioned methods. Most methods use SVM as a classifier while there is no set choice for features and datasets. It can also be observed that the higher the number of images in the dataset lower the accuracy and vice versa.

In summary, a large dataset is not required for hand-crafted methods. These techniques use a few hundred or a few thousand images to train classifiers. Almost 75% of articles discussed in this survey utilized an SVM classifier to classify images into high and low aesthetic levels, and around 15% used support vector regression. Here, the regression provides a continuous score on which threshold is applied for classification into different aesthetic levels. Hand-tuned

**TABLE 4.** Comparative analysis of hand-crafted techniques using content-based features for Bi-level image aesthetic assessment categorization task.

| Authors | Year | Features | | | | | | Classifiers | | | Dataset | | | | | | | No. of Images | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Bag-of-Words | Fish Vector | Dark Channel | SCL | Bag-of-Aesthetics | Global Features | SVM | Adaboost | GMM | DPChallenge | Photo.net | CHUK | CUHKPQ | AVA | Flicker | CHAED | | |
| Nishiyama et al. [30] | 2012 | ✓ | | | | | | ✓ | | | ✓ | | | | | | | 124664 | 77.6% |
| Marchesotti et al. [31] | 2011 | ✓ | ✓ | | | | | ✓ | | | | ✓ | ✓ | | | | | 15581 | 78.0% |
| Tang et al. [32] | 2013 | | | ✓ | | | | ✓ | | | | | | ✓ | | | | 17673 | 83.0% |
| Zhang et al. [33] | 2014 | | | | ✓ | | | | | ✓ | | ✓ | ✓ | | ✓ | | | 40581 | 85.5% |
| Su et al. [34] | 2011 | | | | | ✓ | | | ✓ | | ✓ | | | | | ✓ | | 3k | 92.1% |
| Sun et al. [35] | 2015 | | | | | | ✓ | ✓ | | | | | | | | | ✓ | - | - |

**TABLE 5.** Comparative analysis of deep learning techniques for image aesthetic assessment.

| Authors | Year | Layers | | | Models | Dataset | | | | | | | | | No. of Images | Classification Level | | | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Convolutional | Pooling | Fully Connected | Learning Model & Backbone | Dattra | Photo.net | ATA | Places205 GoogleLeNet | AVA | CUHKPQ | CHUK | LIVE-IQ | Lomas | | Bilevel | Multilevel | High-Low | |
| Lu et al. [88] | 2014 | 6 | 6 | 3 | | | | | | ✓ | | | | | 255k | ✓ | | | 71.20% |
| Zhou et al. [100] | 2015 | 2 | 2 | 1 | | ✓ | | | | | | | | | 29k | ✓ | | | 82.10% |
| Kao et al. [102] | 2016 | 5 | 3 | 1 | | | ✓ | ✓ | | | | | | | 275k | ✓ | | | 79.08% |
| Mai et al. [89] | 2016 | 12 | 5 | 3 | | | | | ✓ | ✓ | | | | | 255k | ✓ | | | 77.10% |
| Wang et al. [90] | 2016 | 5 | 5 | 3 | | | | | | ✓ | ✓ | | | | 273k | | | ✓ | 91.93% |
| Liu et al. [92] | 2018 | | | | Semi-Supervised & Active Learning | | ✓ | | | ✓ | | ✓ | ✓ | | 12k, 3581 & 250k, 779 | ✓ | | | 94.65% |
| Fu et al. [101] | 2018 | | | | Different deep & Learning models | | | | | ✓ | | | | | 250k | | | ✓ | 90.01% |
| Li et al. [85] | 2019 | | | | DenseNet121 | | | | | ✓ | | | | | 250k | ✓ | | | 81.50% |
| Chen et al. [118] | 2020 | | | | ResNet-50 | | | | | ✓ | | | | | 250k | ✓ | | | 83.24% |
| Li et al. [87] | 2020 | | | | Siamese Network | | | | | ✓ | | | | | 250k | ✓ | | | 83.70% |
| McCormack & Lomas et al. [93] | 2021 | | | | ResNet-50 | | | | | | | | | ✓ | 1774 | | ✓ | | 97.00% |
| Zhang et al. [94] | 2021 | | | | InceptionNet | | | | | ✓ | | | | | 250k | ✓ | | | 86.66% |

approaches mainly rely on low-level features and do not consider semantic information of images, providing a minimal scoped aesthetic rating.

### 2) PERFORMANCE OF DEEP LEARNING METHODS

This section presents the comparative analysis of deep learning approaches in terms of layers, learning models, datasets, number of images per dataset, classification level, and accuracy. Table 5 shows the comparison of various deep learning techniques for image aesthetic assessment. Deep learning techniques provide better accuracy than hand-crafted techniques, focusing on the broader picture, including low-level and high-level features. The deep convolutional neural networks require considerable data for training. As the Table depicts, the datasets are more significant than those used in hand-crafted techniques. The depth of the network i.e., the number of layers for each method is also represented in Table 5. Moreover, the accuracy may not be directly proportional to the depth of the network. One should also note that deep learning techniques require more computational resources and time for training and deployment.

## V. LIMITATION AND CHALLENGES

We here list some of the limitations and challenges in the following paragraphs.

- **Lack of Dataset**: The algorithms are trained on various datasets; hence, there is no accurate way to determine the actual performance comparison. The best approach is to fix the dataset for training and evaluation.
- **Open Source Algorithms**: In image aesthetics, most of the algorithms and networks are not open source. The open-source codes are essential for future development and improvement.
- **Lack of Benchmark**: The image aesthetics lack a benchmark dataset to evaluate the algorithms, where each one reports the accuracy on the dataset of their choice. A standard benchmark will help accurately record algorithms' progress in image aesthetics.
- **Parameters Comparison**: The methods in image aesthetics lack comparison on the number of parameters that are critical for many real-time computer vision applications. Unfortunately, existing models only focus on performance without giving any information about the number of parameters and efficiency, which may not be a true representation in the accuracy. Hence, attempts should be made for efficient models for deployment on real-time devices.
- **Generalization**: is a challenging task, and many proposed models only work well on the suggested settings. The mentioned models perform better in one scenario

due to their design for that specific task and fail in other settings. Further, the data can influence the generalization as well as robustness; thus, a significant step is to generalize these algorithms on more generalized tasks.

## VI. CONCLUSION

Images may be degraded due to compression artifacts, illumination or lighting issues, pose or camera angle, sensor problems, background clutter, and other imperfections. Image quality assessment can quantify such degradation. Therefore, image quality can be improved by rectifying degradation in images. With avalanche of digital photographs in major fields of life such as medical & healthcare, information & communication technologies, infotainment, edutainment, and safety & security etc., image quality assessment becomes an essential requirement for decision support. This article presented a comparative study of various image aesthetic assessment techniques covering a wide range of hand-crafted as well as deep learning based models from the year 2005 to 2021. We found that deep learning based models have demonstrated superior performance over hand-crafted based models. Therefore, emerging deep learning based image aesthetic assessment techniques can be incorporated in designing state-of-the-art effective decision support systems for the decision makers of the aforementioned fields.

## CONFLICT OF INTEREST

None declared.

## REFERENCES

[1] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 419–426.
[2] M. Uzair, A. Mahmood, A. Mian, and C. McDonald, "Periocular region-based person identification in the visible, infrared and hyperspectral imagery," *Neurocomputing*, vol. 149, pp. 854–867, Feb. 2015.
[3] A. Chatterjee and O. Vartanian, "Neuroscience of aesthetics," *Annals New York Acad. Sci.*, vol. 1369, no. 1, pp. 172–194, 2016.
[4] 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.
[5] L. Barrett, B. Masquita, K. Ochsner, and J. Gross, "The experience of emotion," *Annu. Rev. Psychol.*, vol. 58, pp. 373–403, Jul. 2007.
[6] H. Maitre, "A review of image quality assessment methods with application to computational photography," *Proc. SPIE*, vol. 9811, pp. 82–96, Dec. 2015.
[7] A. Joy and K. Sreekumar, "Aesthetic quality classification of photographs: A literature survey," *Int. J. Comput. Appl.*, vol. 108, no. 15, pp. 32–36, Dec. 2014.
[8] L. Marchesotti, N. Murray, and F. Perronnin, "Discovering beautiful attributes for aesthetic image analysis," *Int. J. Comput. Vis.*, vol. 113, no. 3, pp. 246–266, Jul. 2015.
[9] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 990–998.
[10] R. Datta, J. Li, and J. Z. Wang, "Algorithmic inferencing of aesthetics and emotion in natural images: An exposition," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 105–108.
[11] M. Rahmani, "A graphical approach for image retrieval based on five layered CNNs model," in *Proc. 5th Int. Joint Conf. Adv. Comput. Intell. (IJCACI)*, 2021, pp. 1–12.
[12] 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.* Cham, Switzerland: Springer, 2016, pp. 662–679.
[13] K.-Y. Lo, K.-H. Liu, and C.-S. Chen, "Intelligent photographing interface with on-device aesthetic quality assessment," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 533–544.
[14] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. 9th Eurpeon Conf. Comput. Vis.*, 2006, pp. 288–301.
[15] M. Redi, N. Rasiwasia, G. Aggarwal, and A. Jaimes, "The beauty of capturing faces: Rating the quality of digital portraits," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 1, May 2015, pp. 1–8.
[16] T. O. Aydın, 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.
[17] E. Mavridaki and V. Mezaris, "A comprehensive aesthetic quality assessment method for natural images using basic rules of photography," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 887–891.
[18] C. Li, A. C. Loui, and T. Chen, "Towards aesthetics: A photo quality assessment and photo selection system," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 827–830.
[19] D. Pogacnik, R. Ravnik, N. Bovcon, and F. Solina, "Evaluating photo aesthetic using machine learning," in *Proc. Slovenian KDD Conf. Data Mining Data Warehouses (SiKDD)*, 2012, pp. 197–200.
[20] W. Yang, "Figure and landscape photo quality assessment based on visual aesthetics," *J. Inf. Comput. Sci.*, vol. 12, no. 7, pp. 2477–2486, May 2015.
[21] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 271–280.
[22] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah, "Towards a comprehensive computational model for aesthetic assessment of videos," in *Proc. 21st ACM Int. Conf. Multimedia*, Oct. 2013, pp. 361–364.
[23] K. Lo, K. Liu, and C. Chen, "Assessment of photo aesthetics with efficiency," in *Proc. 21st Int. Conf. Pattern Recognit.*, 2012, pp. 2186–2189.
[24] 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.
[25] Z. Gao, S. Wang, and Q. Ji, "Multiple aesthetic attribute assessment by exploiting relations among aesthetic attributes," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 575–578.
[26] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.
[27] W. Yin, T. Mei, and C. W. Chen, "Assessing photo quality with geo-context and crowd sourced photos," in *Proc. VCIP*, 2012, pp. 1–6.
[28] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," *Proc. SPIE*, vol. 5666, pp. 149–159, Mar. 2005.
[29] S. Riaz, K. H. Lee, and S.-W. Lee, "Aesthetic score assessment based on generic features in digital photography," in *Proc. AUN/SEED-Net Regional Conf. Inf. Commun. Technol.*, 2012, pp. 76–79.
[30] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 33–40.
[31] 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.
[32] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1930–1943, Dec. 2013.
[33] L. Zhang, Y. Gao, C. Zhang, H. Zhang, Q. Tian, and R. Zimmermann, "Perception-guided multimodal feature fusion for photo aesthetics assessment," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 237–246.
[34] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien, "Scenic photo quality assessment with bag of aesthetics-preserving features," in *Proc. 19th ACM Int. Conf. Multimedia (MM)*, 2011, pp. 1213–1216.
[35] R. Sun, Z. Lian, Y. Tang, and J. Xiao, "Aesthetic visual quality evaluation of Chinese handwritings," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 2510–2516.
[36] A. Mahmood, M. Uzair, and S. Al-Maadeed, "Multi-order statistical descriptors for real-time face recognition and object classification," *IEEE Access*, vol. 6, pp. 12993–13004, 2018.

[37] H. Ullah, M. Uzair, A. Mahmood, M. Ullah, S. D. Khan, and F. A. Cheikh, "Internal emotion classification using EEG signal with sparse discriminative ensemble," *IEEE Access*, vol. 7, pp. 40144–40153, 2019.

[38] 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.*, Jun. 2013, pp. 971–978.

[39] M. Uzair, A. Mahmood, and A. Mian, "Hyperspectral face recognition with spatiospectral information fusion and PLS regression," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1127–1137, Mar. 2015.

[40] M. Ullah, H. Ullah, S. D. Khan, and F. A. Cheikh, "Stacked LSTM network for human activity recognition using smartphone data," in *Proc. 8th Eur. Workshop Vis. Inf. Process. (EUVIP)*, Oct. 2019, pp. 175–180.

[41] T.-F. Lee, P.-J. Chao, H.-M. Ting, L. Chang, Y.-J. Huang, J.-M. Wu, H.-Y. Wang, M.-F. Horng, C.-M. Chang, J.-H. Lan, Y.-Y. Huang, F.-M. Fang, and S. W. Leung, "Using multivariate regression model with least absolute shrinkage and selection operator (LASSO) to predict the incidence of xerostomia after intensity-modulated radiotherapy for head and neck cancer," *PLoS ONE*, vol. 9, no. 2, Feb. 2014, Art. no. e89700.

[42] S. D. Khan, H. Ullah, M. Uzair, M. Ullah, R. Ullah, and F. A. Cheikh, "Disam: Density independent and scale aware model for crowd counting and localization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4474–4478.

[43] *Dpchallenge*. Accessed: Jan. 9, 2018. [Online]. Available: https://www.dpchallenge.com/

[44] J. H. Elder and S. W. Zucker, "Local scale control for edge detection and blur estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 7, pp. 699–716, Jul. 1998.

[45] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *J. Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[46] F. J. J. A. Bianchi, C. J. H. Booij, and T. Tscharntke, "Sustainable pest regulation in agricultural landscapes: A review on landscape composition, biodiversity and natural pest control," *Proc. Roy. Soc. B: Biol. Sci.*, vol. 273, no. 1595, pp. 1715–1727, Jul. 2006.

[47] L. Mai, H. Le, Y. Niu, and F. Liu, "Rule of thirds detection from photograph," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2011, pp. 91–96.

[48] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2206–2213.

[49] 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.

[50] F. R. Al-Osaimi, M. Bennamoun, and A. Mian, "Illumination normalization for color face images," in *Proc. Int. Symp. Vis. Comput.*, 2006, pp. 99–101.

[51] D. Basak, S. Pal, and D. C. Patranabis, "Supprot vector regression," *J. Neural Inf. Process.*, vol. 11, no. 10, pp. 203–224, 2007.

[52] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from Flickr tags," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2007, pp. 103–110.

[53] D. Borth, T. Chen, R. Ji, and S.-F. Chang, "SentiBank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content," in *Proc. 21st ACM Int. Conf. Multimedia*, Oct. 2013, pp. 459–460.

[54] M. Ullah, H. Ullah, and F. A. Cheikh, "Single shot appearance model (SSAM) for multi-target tracking," *Electron. Imag.*, vol. 2019, no. 7, p. 466, 2019.

[55] H. Ullah, I. U. Islam, M. Ullah, M. Afaq, S. D. Khan, and J. Iqbal, "Multi-feature-based crowd video modeling for visual event detection," *Multimedia Syst.*, vol. 27, pp. 1–9, Apr. 2020.

[56] H. Ullah, A. B. Altamimi, M. Uzair, and M. Ullah, "Anomalous entities detection and localization in pedestrian flows," *Neurocomputing*, vol. 290, pp. 74–86, May 2018.

[57] H. Ullah, M. Uzair, M. Ullah, A. Khan, A. Ahmad, and W. Khan, "Density independent hydrodynamics model for crowd coherency detection," *Neurocomputing*, vol. 242, pp. 28–39, Jun. 2017.

[58] H. Ullah, M. Ullah, and M. Uzair, "A hybrid social influence model for pedestrian motion segmentation," *Neural Comput. Appl.*, vol. 31, no. 11, pp. 7317–7333, Nov. 2019.

[59] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang, "Robust late fusion with rank minimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3021–3028.

[60] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah, "Towards a comprehensive computational model foraesthetic assessment of videos," in *Proc. 21st ACM Int. Conf. Multimedia*, Oct. 2013, pp. 361–364.

[61] M. Uzair, R. S. Brinkworth, and A. Finn, "Bio-inspired video enhancement for small moving target detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1232–1244, 2021.

[62] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[63] T.-T. Ng, S.-F. Chang, J. Hsu, and M. Pepeljugoski, "Columbia photographic images and photorealistic computer graphics dataset," Columbia Univ., ADVENT, Boston, MA, USA, Tech. Rep. 205-2004-5, 2005, pp. 205–2004.

[64] D. Lowe, "Object recognitions using local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1150–1157.

[65] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of GIST descriptors for Web-scale image search," in *Proc. ACM Int. Conf. Image Video Retr. (CIVR)*, 2009, pp. 1–8.

[66] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[67] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2390–2398.

[68] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-23, no. 1, pp. 90–93, Jan. 1974.

[69] M. Uzair, A. Mahmood, and A. Mian, "Hyperspectral face recognition using 3D-DCT and partial least squares," in *Proc. Brit. Mach. Vis. Conf.*, vol. 1, 2013, p. 10.

[70] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, 1998, pp. 4–15.

[71] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1473–1476.

[72] M. Uzair and A. Mian, "Blind domain adaptation with augmented extreme learning machine features," *IEEE Trans. Cybern.*, vol. 47, no. 3, pp. 651–660, Mar. 2017.

[73] M. Uzair, F. Shafait, B. Ghanem, and A. Mian, "Representation learning with deep extreme learning machines for efficient image set classification," *Neural Comput. Appl.*, vol. 30, no. 4, pp. 1211–1223, Aug. 2018.

[74] Y. Lifang, Q. Sijun, and Z. Huan, "Feature selection algorithm for hierarchical text classification using Kullback–Leibler divergence," in *Proc. IEEE 2nd Int. Conf. Cloud Comput. Big Data Anal. (ICCCBDA)*, Apr. 2017, pp. 1–4.

[75] R. Schadewald, "Moon and spencer and the small universe," *Evol. J.*, vol. 2, no. 2, pp. 20–22, 1981.

[76] R. Iordache, A. Beghdadi, and P. Viaris de Lesegno, "Pyramidal perceptual filtering using moon and spencer contrast," in *Proc. Int. Conf. Image Process.*, Oct. 2001, pp. 146–149.

[77] K. Liu and J. Yang, "Recognition of people reoccurrences using bag-of-features representation and support vector machine," in *Proc. Chin. Conf. Pattern Recognit.*, Nov. 2009, pp. 1–5.

[78] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, Dec. 2013.

[79] A. Faheema and S. Rakshit, "Feature selection using bag-of-visual-words representation," in *Proc. IEEE 2nd Int. Advance Comput. Conf. (IACC)*, Feb. 2010, pp. 151–156.

[80] G. J. McLachlan and T. Krishnan, *The EM Algorithm Extensions*, vol. 382. Hoboken, NJ, USA: Wiley, 2007.

[81] J. Lewis, "Creation by refinement: A creativity paradigm for gradient descent learning networks," in *Proc. IEEE Int. Conf. Neural Netw.*, Jul. 1988, pp. 229–233.

[82] Y. Takefuji, "Parallel distributed gradient descent and ascent methods," in *Proc. Int. Joint Conf. Neural Netw.*, 1989, p. 584.

[83] J. Xiang and G. Zhu, "Joint face detection and facial expression recognition with MTCNN," in *Proc. 4th Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, Changsha, China, Jul. 2017, pp. 424–427.

[84] Y. Kao, K. Huang, and S. Maybank, "Hierarchical aesthetic quality assessment using deep convolutional neural networks," *Signal Process., Image Commun.*, vol. 47, pp. 500–510, Sep. 2016.

[85] L. Li, H. Zhu, S. Zhao, G. Ding, H. Jiang, and A. Tan, "Personality driven multi-task learning for image aesthetic assessment," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2019, pp. 430–435.

[86] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "Predicting image aesthetics with deep learning," in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst.*, 2016, pp. 117–125.

[87] 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.
[88] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2011–2034, Nov. 2014.
[89] 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.
[90] Y. Wang, Y. Li, and F. Porikli, "Finetuning convolutional neural networks for visual aesthetics," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3554–3559.
[91] Z. Wang, F. Dolcos, D. Beck, S. Chang, and T. S. Huang, "Brain-inspired deep networks for image aesthetics assessment," 2016, *arXiv:1601.04155*.
[92] 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.
[93] J. McCormack and A. Lomas, "Deep learning of individual aesthetics," *Neural Comput. Appl.*, vol. 33, no. 1, pp. 3–17, Jan. 2021.
[94] X. Zhang, X. Gao, L. He, and W. Lu, "MSCAN: Multimodal self-and-collaborative attention network for image aesthetic prediction tasks," *Neurocomputing*, vol. 430, pp. 14–23, Mar. 2021.
[95] K. Bhalla, D. Koundal, S. Bhatia, M. Khalid Imam Rahmani, and M. Tahir, "Fusion of infrared and visible images using fuzzy based Siamese convolutional network," *Comput., Mater. Continua*, vol. 70, no. 3, pp. 5503–5518, 2022.
[96] M. Mahrishi, S. Morwal, A. W. Muzaffar, S. Bhatia, P. Dadheech, and M. K. I. Rahmani, "Video index point detection and extraction framework using custom YoloV4 darknet object detection model," *IEEE Access*, vol. 9, pp. 143378–143391, 2021.
[97] S. D. Khan, H. Ullah, M. Ullah, F. A. Cheikh, and A. Beghdadi, "Dimension invariant model for human head detection," in *Proc. 8th Eur. Workshop Vis. Inf. Process. (EUVIP)*, Oct. 2019, pp. 99–104.
[98] S. D. Khan, H. Ullah, M. Ullah, N. Conci, F. A. Cheikh, and A. Beghdadi, "Person head detection based deep model for people counting in sports videos," in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.
[99] R. Sharma, B. Kaushik, N. Kumar Gondhi, M. Tahir, and M. Khalid Imam Rahmani, "Quantum particle swarm optimization based convolutional neural network for handwritten script recognition," *Comput., Mater. Continua*, vol. 71, no. 3, pp. 5855–5873, 2022.
[100] Y. Zhou, G. Li, and Y. Tan, "Computational aesthetics of photo quality assessment and classification based on artificial neural network with deep learning methods," *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 8, no. 1, pp. 173–282, 2015.
[101] X. Fu, J. Yan, and C. Fan, "Image aesthetics assessment using composite features from off-the-Shelf deep models," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3528–3532.
[102] 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.
[103] A. Giusti, D. C. Ciresan, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Fast image scanning with deep max-pooling convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 4034—4038.
[104] M. Yang, B. Li, H. Fan, and Y. Jiang, "Randomized spatial pooling in deep convolutional networks for scene recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 402–406.
[105] L. Liu, C. Shen, and A. van den Hengel, "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4749–4757.
[106] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nov. 2014, pp. 675–678.
[107] J. Wang, Y. Li, Z. Miao, Y. Xu, and G. Tao, "Euclidean output layer for discriminative feature extraction," in *Proc. 2nd Int. Conf. Image, Vis. Comput. (ICIVC)*, Jun. 2017, pp. 150–153.
[108] S. Tao, T. Zhang, J. Yang, X. Wang, and W. Lu, "Bearing fault diagnosis method based on stacked autoencoder and softmax regression," in *Proc. 34th Chin. Control Conf. (CCC)*, Jul. 2015, pp. 6331–6335.
[109] A. Krizhevsky, T. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[110] I. Melekhov, J. Kannala, and E. Rahtu, "Siamese network features for image matching," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 378–383.
[111] Y. Qi, Y.-Z. Song, H. Zhang, and J. Liu, "Sketch-based image retrieval via Siamese convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2460–2464.
[112] C. Gao, Y. Du, J. Liu, L. Yang, and D. Meng, "A new dataset and evaluation for infrared action recognition," in *Proc. CCF Chin. Conf. Comput. Vis.* Berlin, Germany: Springer, 2015, pp. 302–312.
[113] H. Ullah, S. D. Khan, M. Ullah, F. A. Cheikh, and M. Uzair, "Two stream model for crowd video classification," in *Proc. 8th Eur. Workshop Vis. Inf. Process. (EUVIP)*, Oct. 2019, pp. 93–98.
[114] L. Wang, S. Guo, W. Huang, and Y. Qiao, "Places205-VGGNet models for scene recognition," 2015, *arXiv:1508.01667*.
[115] N. Léonard, S. Waghmare, Y. Wang, and J.-H. Kim, "RNN: Recurrent library for torch," 2015, *arXiv:1511.07889*.
[116] K.-H. Lu, K.-Y. Chang, and C.-S. Chen, "Image aesthetic assessment via deep semantic aggregation," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Dec. 2016, pp. 232–236.
[117] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
[118] Q. Chen, W. Zhang, N. Zhou, P. Lei, Y. Xu, Y. Zheng, and J. Fan, "Adaptive fractional dilated convolution network for image aesthetics assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 14114–14123.
[119] R. Gadde and K. Karlapalem, "Aesthetic guideline driven photography by robots," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1–6.

**ABBAS ANWAR** received the B.S. degree in computer science from Abdul Wali Khan University Mardan (AWKUM), Pakistan, in 2018. He has published in top-tier conferences and journals. His current research interests include machine learning, computer vision, deep learning, image processing, and computing.

**SAIRA KANWAL** received the B.Sc. degree in computer engineering from the University of Engineering and Technology Taxila, Pakistan, and the M.S. degree in electrical engineering from COMSATS University Islamabad, Pakistan. She has more than ten years of experience working as a Computer Vision Engineer at prestigious private and public organizations in Pakistan. Her research interests include computer vision and machine learning.

**MUHAMMAD TAHIR** (Senior Member, IEEE) was born in Peshawar, Khyber Pakhtoonkhwa, Pakistan, in 1981. He received the M.S. degree in computer science from the National University of Computer and Emerging Sciences, Islamabad, in 2009, and the Ph.D. degree in computer science from the Pakistan Institute of Engineering and Applied Sciences, Islamabad, in 2014. From August 2013 to June 2014, he was a Lecturer with the National University of Computer and Emerging Sciences, Peshawar Campus, Pakistan. From July 2014 to July 2015, he was a Lecturer and an Assistant Professor with the City University of Science & Information Technology, Peshawar. He is currently an Associate Professor with the Department of Computer Science, College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia. He has published more than 30 journals and conference articles, two book chapters, and two patents. His research interests include machine learning, pattern recognition, image processing, deep learning, and bioinformatics.

**MUHAMMAD SAQIB** received the Ph.D. degree from the School of Computer Science, University of Technology Sydney, Australia. He is currently a Postdoctoral Research Fellow at the Quantitative Imaging Group, CSIRO-Data61, and a Visiting Scientist at the University of Technology Sydney. Before the current job, he was a Data Engineering Consultant at Virtusa for the Allianz Insurance Data Transformation Project. His research interests include computer vision, pattern recognition, image analysis, data analytic, intelligent transportation, and medical imaging.

**MUHAMMAD UZAIR** received the M.S. degree in electronics and computer engineering from Hanyang University, South Korea, in 2007, and the Ph.D. degree in computer engineering from the University of Western Australia. He worked as an Assistant Professor in electrical engineering at COMSATS University Islamabad, Pakistan, from 2016 to 2018, and later as a Research Associate at the University of South Australia, from 2018 to 2021. He also worked as a Software Engineer at Topcon Precision Systems, Australia. His research interests include machine learning, data analytic, computer vision, and biologically inspired signal processing.

**MOHAMMAD KHALID IMAM RAHMANI** (Senior Member, IEEE) was born in Patherghatti, Kishanganj, Bihar, India, in 1975. He received the B.Sc. (Eng.) degree in computer engineering from Aligarh Muslim University, India, in 1998, the M.Tech. degree from Maharshi Dayanand University, Rohtak, in 2010, and the Ph.D. degree in computer science engineering from Mewar University, India, in 2015. From 1999 to 2006, he was a Lecturer with the Maulana Azad College of Engineering and Technology, Patna. From 2006 to 2008, he was a Lecturer and a Senior Lecturer with the Galgotias College of Engineering and Technology, Greater Noida. From 2010 to 2011, he was an Assistant Professor at GSMVNIET, Palwal. He is currently an Associate Professor with the Department of Computer Science, College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia. He has published more than 60 research papers in journals and conferences of international repute, three book chapters, and holds two patents of innovation, including one U.S. patent. His research interests include Algorithms, the IoT, cryptography, image retrieval, pattern recognition, machine learning, and deep learning.

**HABIB ULLAH** received the M.S. degree in electronics and computer engineering from Hanyang University, South Korea, in 2009, and the Ph.D. degree in information and communication technology from The University of Trento, Italy, in 2015. He worked as an Assistant Professor in electrical engineering at COMSATS University Islamabad, Pakistan, from 2015 to 2016. He worked as an Assistant Professor with the College of Computer Science and Engineering, The University of Ha'il, Saudi Arabia, from 2016 to 2020. He also worked as a Postdoctoral Researcher at the UiT The Arctic University of Norway, in 2020. Currently, he is working as an Associate Professor with the Norwegian University of Life Sciences (NMBU). His research interests include computer vision and machine learning.