Advances and Challenges in Computational Image Aesthetics

Giuseppe Valenzise, Chen Kang, and Frédéric Dufaux

1 Introduction

Decades of advancements in image/video acquisition, coding, and communication have made it possible to capture high-quality pictures and videos using devices within everyone's reach. As a result, a sheer amount of visual data is continuously produced and uploaded to social platforms, e.g., 350 million photos are posted every day on Facebook,[1] and 500 hours of new videos are uploaded on YouTube every minute (as of January 2021).[2] Visual media catalyze and attract people's attention and time, with relevant effects from a social perspective. In particular, they represent an immense ecosystem for marketing, in which the "likes" are the primary source of value (John et al. 2017). In this context, it becomes more and more important to predict in an automatized fashion what a human observer would like to watch, using a computer algorithm. The impact and economic value of such prediction are evident in applications like advertising and communication, personal photo triage, image-based content retrieval, etc. Besides, predicting and understanding what makes up image preference is critical in image enhancement and image recommendation, and, overall, it would contribute to a better understanding of human perception.

[1] https://www.socialreport.com/insights/article/360000094166-The-Latest-Facebook-Statistics-2018.

[2] https://blog.youtube/press/.

G. Valenzise (✉) · C. Kang · F. Dufaux
Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, Gif-sur-Yvette, France
e-mail: giuseppe.valenzise@l2s.centralesupelec.fr; chen.kang@l2s.centralesupelec.fr; frederic.dufaux@l2s.centralesupelec.fr

B. Ionescu et al. (eds.), *Human Perception of Visual Information*,
https://doi.org/10.1007/978-3-030-81465-6_6

The mechanisms underpinning image preference are complex and variegated. In computer science and multimedia, these mechanisms have been studied from different angles including, among others, *interestingness*, *surprise/amazement* and *beauty*. These concepts are often mixed and confused with each other, even if they are clearly associated to different preference processes. Interestingness Gygli et al. (2013) is the ability to attract our attention due to the familiarity of what we know and like. It is produced by either universal factors (popularity of the subject of the image, relevance at a certain historical moment, etc.) or personal factors (link to individuals' life experiences, work, family, tastes, etc.). On the other hand, the *surprise/amazement* mechanism is related to how much the picture content departs from our expectations. Interestingness and amazement are important dimensions to define image *memorability* (Isola et al. 2011), which is the ability to remember the content of the image. Finally, the *beauty* of a picture is the quality or aggregate of qualities that give pleasure to the senses, or pleasurably exalt the mind or spirit (definition from the Merriam-Webster dictionary), and is the matter of study of *aesthetics*. While in the rest of this chapter we will focus on this last mechanism, we stress that all the mentioned processes interact with each other, e.g., image beauty can help predict memorability (Constantin et al. 2019), etc. As a result, it is difficult, if not impossible, to disentangle aesthetic judgments from the other concurrent dimensions. This may introduce significant biases in collecting subjectively annotated datasets targeting one of these specific mechanisms, and represents a considerable challenge in the study of image aesthetics.

In this chapter, we deal with *computational aesthetics* as defined by F. Hoenig, i.e., "the research of computational methods that can make applicable aesthetic decisions in a similar fashion as humans can" (Hoenig 2005). This definition puts the emphasis on both *computability*, i.e., the fact that computational aesthetics should provide measurable output (e.g., a classification as beautiful or not, or a rating on a scale of beauty), and *applicability*, i.e., it should be functional in practical applications. The link between computational and empirical aesthetics lies in the way the human judgments are elicited and collected (which we will discuss further in this chapter when talking about aesthetic datasets). According to Hoenig, computational aesthetics should be restricted to the *form*, and not the content, to make aesthetic computation as objective as possible. However, it is not clear to which extent this separation between content and form can be made in practice, and certainly this difference is not considered in most of the existing aesthetic datasets (which are the essential fuel for modern computational aesthetic techniques).

1.1 What Makes a Picture Beautiful?

Before analyzing computational methods for aesthetic prediction, a natural question that arises is then: *what makes a picture beautiful?* This question has indeed been a matter of philosophical debates for over twenty centuries, and has been closely linked for a long time to the concept of art (at least, for the case of classical Western

arts[3]) (Maître 2018). In ancient Greece and Rome, and in different forms through the Middle Ages and until the Renaissance, aesthetics is dominated by *objectivism*. Beauty is seen as an intrinsic property of an object, which is independent from who looks at it. Classical art implements these universal canons of beauty, which have been coded into well-established rules of proportions, composition, etc. These canons continue to largely inspire art and photography nowadays (e.g., through compositional rules such as the rule of thirds, etc.). This objectivist interpretation provides the foundation to most computational aesthetics methods. On the other hand, *subjectivist* approaches consider beauty as the result of an individual, personal visual experience, summarized by the well-known phrase "beauty is in the eye of the beholder".[4] Subjectivism becomes predominant in the sixteenth century, continuing in romantic and modern art. Among the numerous interpretations of aesthetics, Kant's vision is probably one of the most relevant for computational aesthetics, as it tries to reconcile the subjectivist and objectivist points of view (Zuckert 2007). The universality of beauty is given by "common sense": an object is beautiful not only because it is beautiful for the observer, but also because it is deemed to be beautiful for everybody else. Modern data-driven approaches to aesthetics, which we will discuss later in this chapter, rely somehow on this Kantian interpretation of objectivism, in that they assume aesthetic judgments provided by a pool of human observers approximate the true aesthetic value of a picture.

Modern views on aesthetics tend to agree that objects considered to be "beautiful" have some intrinsic properties recognized by all observers. However, the final decision about whether the object is beautiful or not is purely individual. Neuroscience and experimental psychology seem to support this *interactionist* interpretation: while objective visual cues convey beauty, the resulting aesthetic appraisal is subjective and depends on how the visual cues are processed by higher-level cognitive areas in the brain (Reber et al. 2004). Factors that can affect this processing include cultural background, education, age, mood of the observers, etc. The interactionist viewpoint sets the motivation for a personalized image aesthetics prediction (Park et al. 2017; Ren et al. 2017), where the goal is to adapt a generic aesthetics model for an individual user's preference. We will briefly overview some personalized aesthetic models at the end of this chapter.

Despite the relatively young existence of photography compared to other visual arts, the assessment criteria of pictures have evolved significantly since the first photographic plates in the 1830s. In the early days, photography focused on accurately recording objects, people and scenes (Rosenblum 2008). In the late 1800s, when photography was recognized as an art, photos were assessed using the same criteria as classical paintings. In the twentieth century, several photographic

[3]Notice that this relation has become looser in modern and contemporary art, where producing beautiful depictions is often not the primary purpose of the artwork.

[4]This sentence is attributed to the nineteenth-century Irish novelist Margaret Hungerford. However, the expression has a much older origin, e.g., see Shakespeare's *Love's Labour Lost* (1588): "Beauty is bought by judgment of the eye".

movements started to develop. The realism of photos, which was the most relevant criterion till the beginning of 1900s, was questioned by surrealist photographic movements that developed along with artistic *avant-gardes* of that time. Starting from the 1960s, photography was highly influenced by the development of mass media, advertisement, and pop art, and more recently by digital post-processing, which is nowadays accepted as a part of photographic content creation. As for other forms of art, therefore, the aesthetic assessment of photography is a complex, multi-factorial task, where the influence of the cultural, demographic, and historical contexts plays a crucial role. Thus, it is of paramount importance to specify the scope and objectives of computational aesthetics, which we will discuss in the next section.

This chapter presents an overview of computational aesthetics, including the principal dimensions of analysis, the available sources of annotated data, the algorithmic approaches to predict aesthetic judgments and their performance, as well as the open challenges in the field. We target readers with general knowledge in image processing and machine learning, intending to provide an entry point to this domain through a summary of state of the art, valuable references, and general hints for practitioners and researchers willing to work in this field.

The chapter is organized as follows. We present the main dimensions in computational aesthetics in Sect. 2: this will help us to restrict our attention to general aesthetics, which is the mainstream approach followed nowadays. In Sect. 3, we present some aesthetic datasets proposed in the literature, and we discuss the main aspects to consider when creating or choosing an aesthetic dataset. Section 4 is the core of the chapter and provides a (non-exhaustive) overview of the most popular approaches to predict aesthetics proposed so far, using either hand-crafted or learning-based representations. In Sect. 5, we discuss what we believe are the most urgent challenges in the field of computational aesthetics: dealing with subjectivity, and explaining aesthetic predictions.

2 Dimensions in Computational Aesthetics

There are several dimensions that contribute to creating a taxonomy of image aesthetic quality assessment methodologies, as illustrated in Fig. 1 and discussed below.

2.1 Input Type

Depending on the assumptions made on the type and variety of input images, aesthetic assessment methodologies can be categorized into *general* or *task-specific* methods. The former category aims at predicting the aesthetic value of a picture without making specific assumptions on the content of the image, which can span a broad spectrum of objects and scenes (natural, man-made, portraits, animals, etc.).

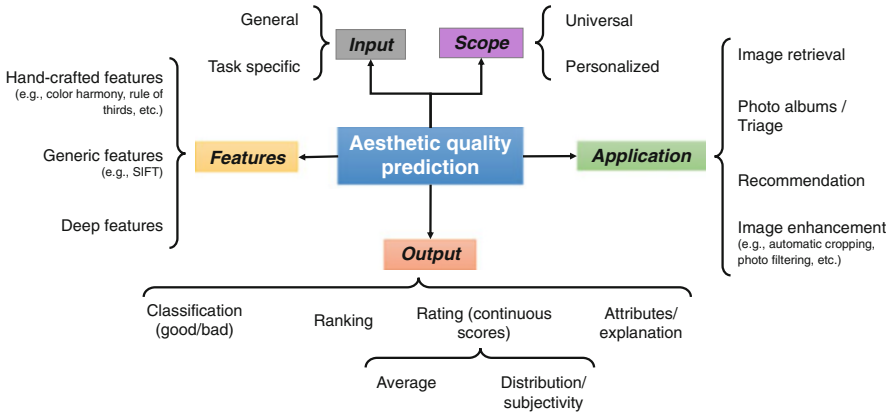


Fig. 1 The different dimensions that compose the aesthetic quality assessment problem

While an a priori knowledge of the semantic content of the picture can greatly aid aesthetic prediction, assuming a closed-set classification setting for image aesthetics would be limiting in some practical applications. Many computational methods proposed in the literature thus do not make this assumption. However, they might internally rely on some form of content classification to improve performance (Luo et al. 2011; Sun et al. 2017). The purpose of a picture can also affect significantly its aesthetic value. For instance, Tifentale and Manovich divide images into several classes (e.g., competitive photography, vernacular, amateur, etc.) and suggest that different evaluation criteria are appropriate for each of them (Tifentale & Manovich 2018). However, most of the existing large-scale aesthetic datasets do not make this distinction. As an example, the AVA dataset (Murray et al. 2012), which is one of the largest reference datasets used in aesthetics, is collected based on photographic challenges but includes as well a large number of amateur-level photographs.

On the other hand, task-specific methods analyze aesthetics for specific kinds of pictures, e.g., aesthetics of faces (Bianco et al. 2018b; Xu et al. 2018), buildings (He et al. 2019), food (Sheng et al. 2018a) or of synthetic images such as video games (Ling et al. 2020). In some particular cases, computational aesthetic approaches can be designed to target non photographic content and artworks, such as paintings (Amirshahi et al. 2013; Hayn-Leichsenring et al. 2017). The general aesthetic problem is more challenging than specific aesthetic tasks, due to the wide variety of content on which minimal or no assumptions can be made beforehand. In the rest of the chapter, we will address the general aesthetic prediction problem, pointing when needed to works addressing specific aesthetic tasks.

2.2 Scope of the Aesthetic Problem

The predictions of a computational aesthetic algorithm can either target a *universal*, "average" observer (or a population of observers), or rather a specific user. In this chapter we mainly discuss the first viewpoint, which is also the most explored in the literature. It is evident that the validity of a universal aesthetic approach is conditioned on the consensus that human observers would achieve in judging the aesthetic value of a picture. Recent methods take into consideration the intrinsic variability in aesthetic assessment across different observers, e.g., they predict a distribution of aesthetic scores or some subjectivity measure (Kang et al. 2019). We discuss the important role of subjectivity in Sect. 5.1.

In contrast with this setting, *personalized* image aesthetics aims to predict the personal preference of a given observer, based on a set of previously annotated pictures or contextual information that enable one to restrict the space of possible aesthetic scores for that person. In this respect, personalized image aesthetics relies substantially on the subjectivist and interactionist foundations of aesthetics. We will briefly discuss personalized aesthetics in Sect. 5.1.3.

2.3 Aesthetic Features

An essential component of any image aesthetic prediction pipeline consists of extracting meaningful features from a picture. The first aesthetic features to be considered were *hand-crafted*, and mainly inspired by guidelines commonly used in photography, such as the rule of thirds, the use of negative space, the color harmony, etc. (see, e.g., Datta et al. (2006), Ke et al. (2006), Luo and Tang (2008), Aydın et al. (2014)), or by mathematical principles, as the classical work of Birkhoff (1933). An advantage of using hand-crafted features is the interpretability of aesthetic predictions. However, the purely objectivist interpretation assumed by these approaches does not take into account the subjective nature of aesthetic judgments, and thus often fails to provide accurate results for a broad range of contents and situations as encountered in real-world applications. We discuss in greater detail hand-crafted methods in Sect. 4.2.

More recently, the availability of large-scale datasets with human annotations (Murray et al. 2012; Kong et al. 2016) has promoted the adoption of data-driven methods, which rely on features extracted from images without a direct association to specific aesthetic attributes or rules. We can broadly consider two classes of features in this category: on one hand, generic features that could be used for other tasks not related to aesthetics (e.g., SIFT (Marchesotti et al. 2011)), and *deep* features learned directly from data. Differently from hand-crafted features, methods based on data-driven features do not look for the presence of specific attributes in the picture, but rather try to infer a relation between image pixels and aesthetic judgments given by humans, which provide the ground-truth for the

evaluation. In this respect, they are less dependent on the initial hypotheses made on the definition of beauty; however, they incur the risk of overfitting the specific conditions in which the features have been learned (e.g., context and methodology of the subjective evaluation, type of content, or hidden patterns in the data). This constitutes a significant challenge toward understanding the factors explaining the predicted aesthetic scores. We present and analyze some relevant deep-learning-based aesthetics approaches in Sect. 4.4.

2.4 Output Prediction

Computational aesthetic methods can predict *classes* (typically binary such as "good/bad" quality, or "amateur/professional", etc.), *ratings* or *rankings* among images. In addition, an algorithm can also predict specific attributes or additional information that can help explain the subjective score (e.g., Aydın et al. (2014)). The first two output types require a single image as input, while the ranking by definition applies to a set of at least two or more images, with the goal to sort them in order of beauty (Kong et al. 2016; Park et al. 2017). The choice between classification and rating is mainly driven by the dataset used, i.e., whether subjective scores have been collected using a binary or any rating scale (discrete or continuous). In some cases, scores originally obtained on a rating scale are converted into binary classes to employ systems trained for classification, e.g., images with average scores less/higher than 5 on a 10 level scale are tagged as bad/good quality. In general, rating scales can provide a better reliability and discrimination of aesthetic scores compared to binary evaluations (Siahaan et al. 2016).

Since ground-truth aesthetic scores are typically obtained by a pool of voters, they represent samples from a distribution of votes. Traditionally, data-driven methods have been concentrating on predicting point estimates such as the average aesthetic score (Deng et al. 2017; Kao et al. 2015). However, recent work tends to estimate directly distributions of scores (Jin et al. 2016a; Talebi & Milanfar 2018; Jin et al. 2018) or measures of subjectivity (Kang et al. 2019), to explicitly model the variability of aesthetic judgments. We discuss in more detail subjectivity prediction in Sect. 5.1.

2.5 Applications

A dimension of analysis of aesthetic quality prediction includes the target applications. These can be varied and range from recommendation to retrieval and enhancement. Some examples of applications that use automatic aesthetic prediction include automatic image cropping (Guo et al. 2018), color (Deng et al. 2018) and composition enhancement (Zhang et al. 2013), photo filter recommendation (Sun et al. 2017), photo triage and album creation (Chang et al. 2016; Kuzovkin et al.

2017), etc. In the rest of the chapter we do not focus on any specific application scenario, but rather on the prediction methodologies.

3 Visual Aesthetics Datasets

Image datasets with aesthetic quality annotations are fundamental to developing computational methods to predict aesthetic appreciation. With the development of computational aesthetics in the mid 2000s, a number of aesthetic datasets were proposed, with different features and label types, to facilitate the training of classifiers based on hand-crafted features. In the 2010s, the creation of large-scale aesthetic datasets such as AVA has enabled researchers to apply deep-learning approaches to this problem, substantially pushing forward the accuracy of aesthetic prediction. In this section we present a review of some popular aesthetic datasets (see Table 1). Our goal is to offer a critical view of some of the main design criteria and trends in constructing aesthetic datasets. To this end, we organize the presentation by discussing some relevant characteristics that are likely to affect the choice of the most appropriate dataset in a given application scenario and the design of new ones.

3.1 Number of Images and Number of Votes per Image

One of the main features of a dataset is its *size*, i.e., the total number of images. Conventional quality assessment datasets collected in lab environments have a limited size of a few tens or hundreds of stimuli due to the costs and time requirements to perform the subjective test campaigns. Datasets obtained through crowdsourcing, instead, can reach a few thousands of stimuli. Finally, crawling annotations from existing websites allows one to obtain hundreds of thousands or millions of annotated images automatically, at the cost of higher noise and possible data bias. For example, the AVA dataset was obtained by crawling over 250k images from DPChallenge (see Sect. 3.2), with an average of 210 votes per image, enabling the use of deep-learning-based methods and becoming a reference dataset in computational aesthetics. We report some statistics of the AVA dataset in Fig. 2.

Often, the total number of votes that can be collected is limited due to time or budget constraints. This is also the case, e.g., of crowdsourcing or lab experiments. In these scenarios, there is a trade-off between the dataset size and the *number of votes per image*. A larger number of images enables better coverage of the vast spectrum of content variety encountered in practical situations. On the other hand, having more votes per image generally yields a better estimation of the picture's aesthetic value, as it reduces the confidence intervals of the estimated scores or score distributions. In technical quality assessment, it is generally recommended

Table 1 Overview of some popular aesthetic datasets according to several characteristics. ACR: Absolute Category Rating; AFC: Alternative Forced Choice pairwise comparison; MOS: Mean Opinion Score

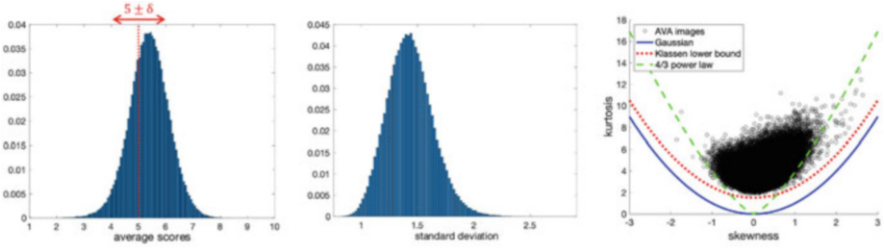
Dataset	Year	Number of images	Votes/image	Image source	Labels	Voting scale	Collection method	Additional labels/attributes
Photo.net (Datta et al. 2006; Datta & Wang 2010)	2006/2008	~ 20k	≥ 10	Photo.net	binary (high/low quality);rating (1-100)	Discrete 1–7	Crawling	"originality", number of views and ratings
CUHK (Ke et al. 2006)	2006	~ 12k	≥ 100	DPChallenge	Binary (high/low quality)	Discrete 1–10	Crawling	N.A.
CUHKPQ (Tang et al. 2013)	2013	17,673	10	Professional photography websites	Binary (high/low quality)	Ternary (low, high, uncertain)	Crawling	7 semantic classes
Hidden Beauty (Schifanella et al. 2015)	2015	~ 15k	≥ 5	Flickr	5-levels discrete ACR	5-levels discrete ACR	crowdsourcing (crowdflower)	4 semantic classes
AVA (Murray et al. 2012)	2012	~ 255k	Between 78 and 549 (avg. 210)	DPChallenge	Discrete 1–10	Discrete 1–10	Crawling	Challenge information, semantic and style labels (for some images)
IAD (Lu et al. 2015b)	2015	1.5M	N.A.	DPChallenge, Photo.net	Binary	Discrete 1–10 and 1–7	Crawling	Camera parameters for some images
AVA-PD (Kairanbay et al. 2019)	2019	~ 119k	..... Same as AVA					AVA + age, gender, location attributes

(continued)

Table 1 (continued)

Dataset	Year	Number of images	Votes/image	Image source	Labels	Voting scale	Collection method	Additional labels/attributes
AVA-reviews (Wang et al. 2019)	2019	40k	..... Same as AVA					AVA + text comments (6 per image)
AVA-Captions (Ghosal et al. 2019)	2020	∼ 230k	..... Same as AVA					AVA + filtered text comments (∼ 5.58 per image)
FACD (Sun et al. 2017)	2017	1280 reference, 28,160 filtered (22 filters per image)	3 comparisons for each filtered image	AVA (8 most popular categories)	Preferences, scores, top preferred filters for each reference	3 AFC	crowdsourcing (AMT)	Semantic classes, filters
Princeton Adobe Photo Triage (Chang et al. 2016)	2016	15,545 (in 5,953 series)	≥ 10 per image pair	User-generated (from personal albums)	Raw preferences, ranking, Bradley-Terry scores	2 AFC + comments	crowdsourcing	Positive/negative comments. categories
AROD (Schwarz et al. 2018)	2018	380k	6,868 on average	Flickr (2k spatial resolution images)	Continuous in [0, 1]	Indirect ("faves")	Crawling	N.A.
PCCD (Chang et al. 2017)	2017	4235	N.A.	gurushots.com	Text comments, normalized rating	Rating scale 1–10, text comments	Crawling	7 aesthetic attributes
AADB (Kong et al. 2016)	2016	10k	5	Flickr	Score distribution, attributes	5-levels discrete, positive/negative attributes	crowdsourcing (AMT)	11 attributes, individual rater IDs

IAE (Yu et al. 2019)	2019	22k	10	Flickr and Instagram	Ratings, binary classes	4-level ACR	lab (aesthetics), crowdsourcing (emotions)	Emotion categories
Waterloo IAA (Liu & Wang 2017)	2017	1k	26	Photo.net	MOS	Single stimulus continuous	lab (ITU rec.)	5 semantic classes
FLICKR-AES (Ren et al. 2017)	2017	40k	5	Flickr	Ratings	5-level discrete ACR	crowdsourcing (AMT)	rater ID
REAL-CUR (Ren et al. 2017)	2017	~ 2870	1 (with repetitions)	Personal albums	..... N.A.			Rater ID
EVA (Kang et al. 2020)	2020	4070	≥ 30	AVA (medium-high quality)	Aesthetic scores, attribute scores, attribute importance	11-levels ACR (global score), 4-level Likert scale (attributes), binary (attribute importance)	crowdsourcing (custom website)	raters ID, voting time, voting difficulty, 6 semantic classes



(a) Distribution of avg scores (b) Distribution of std (c) Skewness-Kurtosis maps

Fig. 2 Some statistics of the AVA dataset (Murray et al. 2012), perhaps the most popular dataset used in computational aesthetics. (**a**) Normalized distribution of the average scores of each image. The distribution can be modeled by as a Gaussian, with an average of 5.38, which is slightly lower than the mid-point of the rating scale (i.e., 5.5). Many computational aesthetics methods obtain binary labels from these scores, by labeling as *high-quality* those images with scores larger than $5 + \delta$, and as *low-quality* those images with scores lower than $5 - \delta$. The images with average scores in the interval $5 \pm \delta$ are often discarded as they are considered aesthetically ambiguous. Notice that this interval is *not* symmetric around the mean score of the dataset. (**b**) Normalized distribution of the standard deviations of the image scores. It has a longer tail (images with high std) compared to a Gaussian. (**c**) Skewness-Kurtosis maps (Park & Zhang 2015) can be used to visualize the consensus in the scores, and can be matched against theoretical bounds (here, the bound for a truncated Gaussian distribution; the Klassen lower bound for unimodal distributions; and a power law). See Sect. 5.1 for further details on the interpretation of these maps

that stimuli are voted by at least 15 observers (ITU-R 2012), with the underlying assumption that the distribution of votes is unimodal and approximately normal. This is not often the case for aesthetic quality assessment, where score distributions could be multimodal or strongly skewed, and thus a higher number of samples might be necessary. Furthermore, in lab experiments, all the stimuli are generally voted by the same set of raters (allowing one to apply some inter-rater agreement reliability analysis (Siahaan et al. 2016)), which is rarely the case for large-size datasets.

The trade-off between dataset size and score precision on video quality prediction using a deep neural network has been investigated in Götz-Hahn et al. (2019). Interestingly, the authors find that, when the total budget of votes is sufficiently high (larger than 1000 votes), the quality prediction performance appears relatively stable. For example, for a total budget of 100k votes, training prediction models based on deep neural networks using 1000 images with 100 votes per image, or 100,000 images with only one vote per image, produces quality score predictions with similar accuracy. Conversely, for smaller budgets (of 1000 images or less), intermediate budget allocations (e.g., five votes for 200 different images) provide higher performance. Notice that the quality evaluation task in Götz-Hahn et al. (2019) targeted technical video quality as intended in a video streaming setting rather than aesthetics. An extension of these observations to aesthetic quality is still missing.

3.2 Image Source

Depending on their source website or device, the images in a dataset might have very different technical and aesthetic qualities. Similarly, their annotations could vary significantly across data sources, e.g., they can be given by people with little background or knowledge in photography, groups of knowledgeable practitioners, or even professional photographers. A typical source of annotated images is photo amateurs and professional websites, such as Flickr, Photo.net, DPChallenge, etc., and social media platforms such as Instagram.

Flickr is probably the largest public source of photos online, with several hundreds of billions pictures hosted by the website. The uploaded pictures come with a number of metadata, including photographic attributes such as exposure time, aperture, camera model, and in some cases geolocalization. In addition, for each image it is possible to get the number of views and the "faves", i.e., the number of times an image has been liked by users. This information is used in some works as a proxy to aesthetic scores (Schwarz et al. 2018).

Photo.net is one of the oldest photo repositories used to produce aesthetic datasets. It hosts almost 5 million high-quality pictures taken by photographers with different experience from hobbyists to professionals. Datta et al. (2006), Datta and Wang (2010) collected one of the first aesthetic datasets based on Photo.net, which has been thereafter referred to with the same name as the website. Images from Photo.net have two kinds of annotations: aesthetics and originality, both rated on a discrete scale with 7 levels. Later versions of the website fused the two attributes in a single value, based on the observation that the two quantities are highly correlated.

DPChallenge is another website for photography amateurs and enthusiasts, which collects over 650k images organized in more than 3000 weekly thematic contests (challenges). The challenges are a fundamental component of the website to motivate users to submit their pictures, which span a broad range of qualities. Each photo can be voted on a discrete scale with 10 levels. The distribution of the average image scores is well modeled as a normal distribution with an average slightly higher than 5, while the standard deviation of the scores is slightly positively skewed with a longer tail. The number of votes per image can be significant (in the order of several hundreds). However, the aesthetic scores can be highly influenced by the thematic context of the challenge. The same holds for the subjectivity of the collected scores (Kang et al. 2019). The popular AVA dataset (Murray et al. 2012) has been created from DPChallenge, and is itself often used as a source to build other aesthetic datasets (Kairanbay et al. 2019; Kang et al. 2020). DPChallenge has inspired more recently other websites such as *Gurushots.com* and *500px*, which also employ similar concepts as the challenges, and collect user comments with the goal to offer personalized advice and improvement tips to photographers.

Finally, some datasets do not rely on online resources to collected voted images (e.g., to avoid copyright issues), and rather employ personal pictures or photo albums (Chang et al. 2016).

3.3 Voting Methodology and Aesthetic Labels

Existing datasets have been collected with different methodologies and experimental procedures, which makes it difficult in general to compare aesthetic scores across databases. Siahaan et al. (2016) have studied the impact of the voting scale on the reliability and repeatability of subjective aesthetic scores. They find that a 5-level absolute category rating (ACR) scale provides mean opinion scores (MOS) with better *reliability* (which can be measured, e.g., by inter-observer agreement) and *repeatability* across different datasets. Other rating scales, and in particular categorical binary scales (e.g., "high/low quality") tend to produce noisier aesthetic labels and thus are not recommended. Unfortunately, a large part of the datasets available in the literature seems not to respect these recommendations.

The choice of the questions and adjectives in the voting scale is critical in aesthetics. Differently from conventional technical quality assessment (ITU-R 2012), only few datasets employ some form of training of the raters to ensure that the task is clear and to provide examples of the stimuli used in the test (Kang et al. 2020; Schifanella et al. 2015; Liu & Wang 2017). Pairwise comparisons approaches can partially solve this issue, as they require choosing the preferred stimulus between two alternatives (two-alternative forced choice, or three-alternative forced choise in case a tie option is given). Pairwise comparisons involve a smaller cognitive load, and eliminate the need for training. However, the number of pairs to compare grows quadratically with the number of stimuli, which requires in practice the use of some form of approximate design (Li et al. 2013) or active sampling (e.g., Ye et al. (2014)). The collected preferences can be transformed into relative quality scores by applying some heuristics (e.g., vote counts) or psychometric scaling (Chang et al. 2016), such as the Thurstone or the Bradley-Terry-Luce (BTL) models. Fusing rating scales and pairwise preferences, e.g., to merge or align subjective datasets, is an active research topic (Zerman et al. 2018; Perez-Ortiz et al. 2019), which is still unexplored for aesthetics.

The labels made available in aesthetic datasets may include the simple raw data, or some form of processed data. In the CUHK dataset (Ke et al. 2006), for instance, the average rating scores are filtered to remove images with uncertain quality (those lying in the middle of the rating distribution), and only the top/bottom 10% of the pictures are retained and classified as high/low quality. A similar strategy is typically followed to create binary labels for classification on the AVA dataset (Murray et al. 2012), by discarding images with an average score between $5 - \delta$ and $5 + \delta$ (with $\delta = 0$ corresponding to using the whole dataset, see Fig. 2a). Typical values of δ range between 0 and 2.5.

In some cases, the raw scores are collected in an *indirect* way, by retrieving different but presumably related information, and require further processing to be converted into aesthetic labels. For example, the authors of Suchecki and Trzciski (2017) collect 1.7 million photos from Flickr, and assign them an aesthetic score which is a function of the average number of daily views of the picture. The AROD dataset (Schwarz et al. 2018) also crawls images from Flickr but considers the

number of "faves" in the equation. While this data is largely available and cheap to collect, "faves" or "likes" are only loosely connected to aesthetics, and might be rather related to other preference mechanisms (interestingness, amazement), as discussed in Sect. 1.

3.4 Collection Method

There are essentially three approaches to collect aesthetic annotations. In *laboratory* experiments, the pictures are voted by a pool of observers in a particular test room, typically illuminated and equipped according to quality assessment recommendations such as the ITU-R BT.500 (ITU-R 2012) to provide controlled and reproducible testing conditions. Lab experiments generally include a subject screening for visual acuity/color perception, and a training phase, which depends on the methodology, to present the rating scale, the nature of the quality attribute to evaluate, and the use of the voting interface. Subjective quality campaigns performed in the labs are generally the best option to obtain precise and reliable subjective scores. However, they entail a significant cost in terms of data collection time—the use of a special test environment makes it impossible to massively parallelize the test.

Crowdsourcing resolves the limitations of lab experiments, in that they enable massive parallel voting, at the cost of reliability and repeatability. These are inevitably degraded due to the lack of effective controls of the engagement of raters, as well as the huge variety in the display devices, internet connection quality and viewing conditions. To partially alleviate this problem, it is highly advisable to include quality checks (such as "gold standards" test questions) in such a way to enable later the detection and filtering of potential unreliable votes or raters. Examples of quality checks for aesthetic crowsourcing are available, e.g., in Schifanella et al. (2015), Siahaan et al. (2016), Chang et al. (2016). Crowdsourcing has become one of the most popular approaches to collecting subjective scores (see, e.g., Ribeiro et al. (2011)), and has been employed in many aesthetic datasets.

Finally, a common approach that has been used to build aesthetic datasets consists of *crawling* aesthetic annotations (ratings, comments, preferences) directly from existing online sources, as described in Sect. 3.2.

3.5 Additional Labels and Attributes

In addition to aesthetic scores, datasets can offer additional labels to enable multi-task applications (Kao et al. 2017b), or provide contextual information for aesthetic prediction. Typical additional labels include the semantic class of the picture, generally categorized based on the content, e.g., nature, portraits, buildings, etc. In some

cases, the aesthetic data is complemented by textual annotations and comments crawled from the web or collected during the experiments. The text information has been used to provide aesthetic explanations, leveraging natural language processing architectures (Wang et al. 2019). Perceptual attributes directly contribute to aesthetic judgments, and some datasets focus on measuring them, although not in an aesthetic context. It is the case, for example, for colorfulness (Zerman et al. 2019) or dynamic range (Hulusic et al. 2016). Other datasets provide additional attributes such as the emotional response, which are not directly related to aesthetics, but can participate in image preference formation (Yu et al. 2019). Finally, aesthetic scores can be augmented with unique identifiers of voters, to facilitate personalized aesthetics applications.

4 Approaches to Computational Aesthetics

In the following, we review the main approaches to computational aesthetics proposed in the literature. Two general families of methods can be distinguished: those based on hand-crafted or generic features, and those that try to deduce the aesthetic quality of a picture directly from data, in an end-to-end fashion. Before presenting in more details these two paradigms, we briefly describe some preliminary work aimed at defining a mathematical model of aesthetics. All the methods introduced here build on an objectivist interpretation of aesthetics. Readers interested in computational aesthetics can also refer to the experimental survey of Deng et al. (2017).

4.1 Mathematical Approaches

Although it does not explicitly provide an algorithm to compute aesthetics on a computer (in fact, computers had not yet been invented at that time), the mathematical theory proposed by the mathematician and statistician George D. Birkhoff in 1933 (Birkhoff 1933) is generally considered as the predecessor of all quantitative models of aesthetics. Formalizing the artistic principle of "unit in variety", Birkhoff suggested the measurement of aesthetics as a ratio:

$$M = \frac{\text{Order}}{\text{Complexity}}. \tag{1}$$

The aesthetic measure can then be interpreted as the reward that the observer gets in terms of perceiving a pleasing harmony (order) when putting in an effort to focus and integrate a scene (complexity).

Despite his efforts to prove the validity of his conjecture in different fields of arts, Birkhoff was not able to bring convincing empirical evidence to his theory, also due

to the lack of modern mathematical and signal processing tools to analyze pictures. Nevertheless, Birkhoff's ideas have been rediscovered and utilized in later work, with the aid of more modern mathematical tools, e.g., the Kolmogorov complexity is employed in Machado and Cardoso (1998) and Rigau et al. (2008) to compute the complexity of the image (a JPEG or fractal compression of the picture are used to approximate the Kolmogorov complexity, which is not computable), together with more sophisticated image processing tools such as image segmentation. Recently, a mathematical formulation of aesthetics based on thermodynamics that partially extends the principles of Birkhoff has been proposed in Lakhal et al. (2020).

4.2 Hand-Crafted Features

Modern approaches to computational aesthetics have abandoned the search for a holistic mathematical formulation of beauty in favor of a more pragmatical data-driven vision of the problem. The hypothesis is that aesthetics resides in a set of attributes and features of an image, and the relation between these features and aesthetic judgment can be deduced by observing a large number of pictures annotated by humans. The general pipeline of this data-driven approach consists therefore of three steps: (1) choose or collect a photographic dataset with aesthetic annotations; (2) extract a set of relevant image features from each photo in the dataset; (3) train a classifier (typically, a support vector machine—SVM) or a regressor to predict aesthetic scores based on the extracted features of unseen images (Kuzovkin 2019). By relevant features, we intend features that can be related to specific aesthetic attributes (color, composition, etc., see Fig. 3 for some examples). These features provide valuable information to the classifier or regressor, which learns how to combine them to produce a synthetic overall aesthetic score. Since we already discussed the collection of aesthetics datasets in Sect. 3, we will focus on the feature extraction and the prediction scheme in the following section.

4.2.1 Initial Works

Two seminal works in modern computational aesthetics were proposed by Datta et al. (2006) and Ke et al. (2006) in 2006. In addition to collecting the first aesthetics datasets, they introduce a set of aesthetic features and a general prediction framework based on classification (e.g., using a support vector machine—SVM) to determine if a picture has a high or low aesthetic level. Many later works follow a similar approach and use similar features.

Datta et al. (2006) collected the Photo.net dataset, containing approximately 3800 pictures (see Table 1). They consider 56 features, including:

- *low-level and color* features such as the average pixel intensity to characterize the use of light (exposure); a colorfulness measure computed as a distance between

Fig. 3 Some photographic rules and concepts that serve as models to design aesthetic features. (**a**) The rule of thirds is a well-known composition rule suggesting that salient objects in the picture should be positioned along or at the intersections ("powerpoints") of the horizontal/vertical lines dividing the height and length of the image into 3 equal parts. (**b**) Negative space is the area surrounding the main subject in the photo (positive space), which should be left unoccupied to facilitate the focus of the observer on the region of interest. A disregard for negative space may produce cluttered and unclear pictures. (**c**) The depth of field is the distance between the closest and farthest objects in a photo that appear sharp. Using a low depth of field (an effect sometimes referred to as *bokeh*) is a powerful way to concentrate the attention on the subject of the picture (by emphasizing the negative space through blur), and is considered aesthetically appealing. (**d**) Similar to harmony in music, colors in photography can produce more or less harmonic combinations. The rules of color harmony are numerous (see, e.g., Moon and Spencer (1944)). They are based on the principle of avoiding colors that are too close on the color wheel (shown in the right part of the image), which would create ambiguity (similar to dissonance in music). Instead, an aesthetically pleasing combination should include complementary colors or combinations of colors lying on simple geometric shapes on the color wheel (e.g., in this example, the three main colors can be imagined to be at the vertices of a triangle). Figure best viewed in color

the distribution of color (in the LUV color space) of the image and a reference distribution with uniform color probabilities; the average saturation and hue;

- *composition*-related features, which are inspired by photographic rules. These include a measure of the *rule of thirds*, computed as the average intensities in the center portion of the image, in the HSV color space; an indicator of the *depth*

of field based on a wavelet decomposition of the image; aspect ratio; a region composition indicator based on color segmentation;
- *familiarity*, intended as the average distance of an image to other images in the dataset in terms of color, texture and shape;
- *texture* features based on a wavelet decomposition in the HSV space to quantify the graininess or smoothness of the textures;
- *shape convexity* features, which compute the portion of the image containing convex objects, and are related to the assumption made by authors that convex and regular shapes produce a positive aesthetic response.

An SVM classifier trained with a selected subset of these features obtains an accuracy ranging between 62 and 70%, depending on the margin left between the ground-truth binary classes. This system has been later extended in Datta and Wang (2010) and has been put online with the name ACQUINE (aesthetic quality inference engine), which computes an aesthetic rating for a given input image.

The work of Ke et al. (2006) has a similar approach. The goal is to classify whether an image is a professional or amateur picture. To this end, the authors crawled 60,000 photos from DPChallenge, choosing the ones voted by at least one hundred viewers. The two aesthetic classes are obtained by taking the highest and lowest 10% average rates. The features proposed in this work try to capture mainly high-level photographic concepts by using image processing and computer vision tools, and include:

- two *simplicity* measures. One is computed from edge maps in the picture: in professional pictures the edges are concentrated round the middle of the image, reducing the quantity of distracting structure in the background (similar to the concept of negative space in photography, see Fig. 3); the other is the hue count, another way to gauge the cluttering of a photo;
- *color palette*, computed as the histogram of a version of the image with quantized color levels. The number of professional/amateur photos that are the nearest neighbors to the current image in this histogram space determine the class of the picture;
- *low-level features*, including a measure of blur, and intensity features such as contrast and exposure.

These features are then used into a naive Bayes classifier to discriminate between professional and amateur photos. The reported classification accuracy peaks at 72% when professional/amateur photos correspond to the 10% highest/lowest average scores. Later work show that for less favorable class splits, the accuracy is lower and generally ranging between 60 and 70%.

4.2.2 Considering the Salient Object of the Picture

The two methods discussed above obtained encouraging performances, although the accuracy is still relatively limited. Later work has further improved classification

accuracy by extending the feature set and/or the classification strategy. A class of methods takes in consideration explicitly the role of the *subject* of the picture. For example, Luo and Tang (2008) employ a similar approach as Ke et al. (2006), but they compute different criteria depending on whether an image region belongs to the subject or to the background. The distinction subject/background is done based on a simple blur-based heuristic. Mai et al. (2011, 2012) analyze the salient regions of an image using a saliency map predictor, to determine whether the composition of the photo respects the rule of thirds and the principle of simplicity (e.g., by using the negative space or a low depth of field, see Fig. 3). Zhang et al. (2014) adopt a more sophisticated approach inspired by human perception, where aesthetics is evaluated along *visual scan paths* (represented as graphlets), to mimic human visual attention mechanisms. The idea to embed visual attention mechanisms in computational aesthetics has been further explored with deep-learning-based methods (see Sect. 4.4).

4.2.3 Including Semantic Information

Another strategy to augment aesthetic features consists in taking into account the semantics of the picture, and in particular high-level features related to the *image content*. For instance, Dhar et al. (2011) employ a complex set of features, including both low-level ones (as in the works described above) and high-level features describing composition (depth of field, salient object, etc.), content (faces, presence of animals, indoor-outdoor, etc.) and sky illumination. The high-level descriptors are obtained by several classification subsystems (SVM classifiers), a scheme that scales poorly with the number of possible objects to recognize. As we will see next, this limitation is partially solved by using deep learning models, which can easily represent and predict a vast ensemble of object classes. Image content significantly affects which visual features are relevant to predict aesthetics (e.g., the way to perceive beauty of a landscape is forcibly different from the aesthetics of portraits) (Simond et al. 2015). In this respect, Luo et al. (2011) mix the subject detection strategy with image categorization and propose a different subject/background segmentation and extract visual features differently depending on the class of the picture.

4.2.4 Multi-Dimensional Approaches

Some methods based on hand-crafted features do not simply aim at predicting a global aesthetic class or score, but rather treat aesthetics as a *multi-dimensional* problem, where the overall evaluation is obtained as the composition of several aesthetic attributes. This viewpoint has the advantage to provide a better interpretability of *why* an image is aesthetically pleasing or not. Lo et al. (2012) propose a visual interface with a sort of "radar" plot (see Fig. 4) where the magnitude of five attributes (saturation, color, composition, contrast and richness) is displayed. The

(a) original (b) edited (c) aesthetic attributes

Fig. 4 The multi-dimensional representation of aesthetics proposed in Aydın et al. (2014). For each image, five photographic attributes are evaluated. The overall aesthetic score is given by a combination of the attribute scores. Decomposing the aesthetic scores into multiple components enables one to explain why a photo is aesthetically pleasing, and can be used to guide an enhancement process. In this example, an original image (**a**) with low dynamic range (tone) and drab colors is edited to increase colorfulness and contrast, while also putting more emphasis on the subject (**b**). The attributes scores for the two images can be intuitively displayed in a radar plot (**c**). The area enclosed by the polygon in the plot gives an indication of the overall aesthetic score. Figure best viewed in color

surface of the polygon connecting the different attribute scores give an indication of the overall aesthetic quality. A similar approach is proposed in Aydın et al. (2014), where the attributes are linked to photographic concepts and are calibrated by an original experimental procedure. On the opposite of these multi-dimensional approaches are methods that consider aesthetics from the perspective of a single attribute, e.g., by considering only color harmony (Lu et al. 2015a, 2016).

4.2.5 Leveraging Users' Comments

In addition to visual features, some datasets report also text comments from users (see Sect. 3). This data can provide valuable information to predict aesthetics. For example, the authors of San Pedro et al. (2012) employed hidden Markov models to analyze text comments crawled from DPChallenge. They compared the features associated to text with image-based features (combined using a support vector regression to predict aesthetic scores), and found that, interestingly, the text-based features perform substantially better than image-based ones on a regression task. The fusion of text and image features provide only a marginal advantage. It must be noted, though, that the feature extraction mechanism for text comments is likely to generalize poorly to comments using expressions not contained in the dataset. We will see next that the idea of employing text comments has been further exploited in the context of deep-learning-based methods, where comments are also generated by the prediction algorithm to endow the aesthetic judgments with partial explainability.

To conclude this overview on hand-crafted approaches, it is worth mentioning works targeting task-specific input (and not general aesthetics as for the methods described above), such as images of people. In those cases, features describe specific

aspects related to faces, such as the pose, the expressions and lighting (Li et al. 2010; Redi et al. 2015).

4.3 Generic Features

So far we have discussed methods that try to encode explicitly the best practices of photography. The advantage of these methods is that, in many cases, it is possible to identify the factors that lead to a certain aesthetic score. However, the performance of hand-crafted features rest limited due to several reasons, e.g., the features are not exhaustive (they cannot cover all the possible photographic principles), and they are based on simple heuristics, i.e., they try to encode complex rules by simple, low-level processing. As a result, these methods have a low ability to generalize to similar cases, resulting in a generally large variance of the prediction performance.

Marchesotti et al. (2011) proposed a very different approach. Instead of using specific aesthetic features, they argue that the aesthetic information is implicitly embedded into *generic* image features, which encode the distribution of local image statistics. The motivation behind this approach is that, at the time this work was proposed, generic image features such as the Bag of Visual Words (BOVW, Csurka et al. (2004)) and Fisher Vectors (FV, Jaakkola et al. (1999)) displayed excellent capabilities to deal with complex semantic tasks, which suggests that they could also lead to good performance for aesthetics. The hypothesis is that generic local features can reveal information about the local sharpness or color distribution that, when aggregated from patch level to image level, is sufficiently rich to summarize the global characteristics of images (mix of sharp and blur edges, color harmony, etc.). In this respect, hand-crafted features capture specific instances of these global characteristics. To test this hypothesis, the authors extract SIFT (Scale-invariant feature transform) features from the image. The SIFT features describe the local gradient orientations at keypoints detected by a scale-space blob detector (Lowe 1999). In addition to SIFT, some color descriptors are also considered. The features are aggregated at the image level, using either a discrete histogram (BOVW), or a more sophisticated modeling of the second-order statistics (FV) using a high-dimensional Gaussian mixture model, which yields continuous features. The two features are inputted to an SVM classifier to predict the aesthetic class (high/low quality). The results obtained by the authors on the Photo.net and the CUHPK datasets (see Sect. 3) show significant gains (from 5 to 10%) in terms of accuracy compared to hand-crafted approaches such as Datta et al. (2006) and Ke et al. (2006).

The results of Marchesotti et al. (2011) are particularly relevant in the field of computational aesthetics, since they demonstrated for the first time that generic, aesthetic-agnostic features could outperform a carefully hand-crafted feature design based on well-established photographic rules. Later, the same authors extended their work to add some form of explainability, by including text comments from AVA and mining them to discover relevant aesthetic attributes (Marchesotti 2013). These

works prelude a trend that has become the main approach in computer vision and multimedia nowadays, i.e., learning generic features directly from data using deep neural networks.

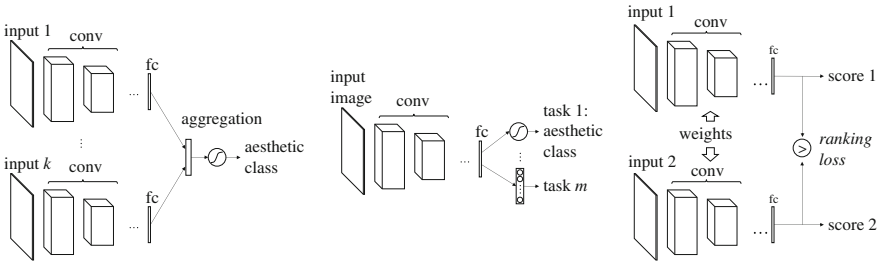
4.4 Deep Learning Approaches

The method based on generic features presented above is still employing a handcrafted design of low-level features (SIFT or color descriptors). In other words, the design of the features is *independent* of the data, and the task of making an efficient use of them to predict aesthetic scores is left to the classifier. The advent of deep neural networks changed significantly the paradigm of feature extraction, making it *data driven*: a high-dimensional (often, in the order of 10^6 parameters) neural network model is learned in an end-to-end fashion, by optimizing a differentiable loss function using directly the images and the corresponding labels as input, without the need to pre-compute any handcrafted features. A class of deep neural networks of particular interest for image processing is convolutional neural networks (CNN). The interested reader can refer to Goodfellow et al. (2016) for an introduction to deep learning. We review in the following some of the main approaches and challenges to employ deep convolutional neural networks for computational aesthetics.

4.4.1 Preserving Global and Local Information

As mentioned above, deep neural networks typically contain millions of parameters to learn (called also weights), e.g., the VGG-16 architecture (Simonyan & Zisserman 2014) used in many aesthetic works has 134 millions of parameters. This makes their use very demanding both in terms of computational time and memory consumption (Bianco et al. 2018a). In practice, to keep the problem tractable with the available graphical processing units (GPUs), especially at the beginning of the deep learning era input images were resized to a lower resolution (e.g., 224×224 pixels) in order to be used on pre-trained models, which were then fine-tuned for a specific application. Nevertheless, resizing images to small, square thumbnails in the case of aesthetic evaluation can seriously alter both the composition of the image and the presence of small but relevant details, compromising aesthetic assessment. Initial works applying CNN architectures to computational aesthetics addressed this issue.

The first deep-learning-based system for aesthetic classification was proposed by Lu et al. (2014) under the name of **RAPID** (Rating pictorial aesthetics using deep learning). To deal with the resizing and aspect ratio problems, RAPID employs a *two-column* network (see Fig. 5a): two identical networks (in this case, AlexNet is used (Krizhevsky et al. 2012)) with independent weights are fed with different inputs, and their features are then merged into one or more shared layers (typically



(a) Multi-column CNN (b) Multi-task CNN (c) Siamese network

Fig. 5 Some deep neural network architectures used in computational aesthetics. (**a**) Multi-column CNNs are a way to handle different inputs (images, attributes, patches, etc.). These are processed by parallel networks, which could have or not the same architecture and shared weights. The output of the columns is then merged in an aggregation layer to obtain an aesthetic class or rating. (**b**) Multi-task networks are designed instead to perform different tasks which are correlated. The input image is processed by a single network, and the tasks are differentiated at the last layers. The difficulty with these networks is to find a good balance between tasks in the training. (**c**) Siamese networks are composed by two identical networks (with shared weights), which are trained simultaneously by minimizing a *ranking loss*

fully connected). The two networks are trained jointly. The first column in RAPID takes as input the whole picture, warped (resized and padded) to 224×224 spatial resolution. In the second column, the input is a patch (again of 224×224 pixels) randomly extracted from the image at the original resolution. The evaluation of the two columns is repeated 50 times to average the results across different random patches. In this way, the network learns to evaluate global and local information, both necessary to predict the aesthetic class of an image. RAPID achieves 73% classification accuracy on the AVA dataset, which is higher than any other previously proposed hand-crafted features on this dataset. The performance is slightly improved (74%) when adding an extra column to the network with style information (available for some images in AVA).

This approach is later extended in Lu et al. (2015c), which proposes a deep multi-patch aggregative network (**DMA-Net**) with five columns. In this case, the input to each column is an original-resolution patch extracted ramdomly from the image, and the five branches are sharing weights to speed-up training. The features from the columns are merged using either an order-independent pooling operator (e.g., average or max pooling), or using a fully connected network with a sorting layer. The reported classification accuracy with the best configuration is 75.4%. A different strategy is considered by Mai et al. (2016) in the multi-net adaptive spatial pooling CNN (**MNA-CNN**). They add an adaptive spatial pooling layer upon the regular convolution and pooling layers to handle a limitation of the conventional CNN design, where the presence of fully-connected layers assumes a fixed-size feature vector. The idea is to perform max pooling over local image regions, but fixing the output size instead of the receptive field's size. This strategy is repeated

for different adaptive spatial pooling sizes to obtain a multi-scale representation. MNA-CNN achieves a classification accuracy of 77.1% on AVA.

The multi-column principle introduced by DMA-Net has proved to be very effective in preserving local and global information, and has been employed by many deep-learning-based approaches later on. Ma et al. (2017) improved the selection of patches in their adaptive a layout-aware multi-patch (**A-Lamp**) CNN. Differently from DMA-Net, A-Lamp selects patches adaptively based on the content of the image, using a pre-trained saliency model. An attribute-graph representation of salient patches is then assembled using the areas of the patches, as well as their reciprocal orientation and distance. This information is processed by layout-aware sub-network to capture the topology and layout of the picture. The selected patches follow then a multi-patch sub-network with an aggregation layer at the end, similar to DMA-Net. Finally, the two subnets are merged through a learned aggregation layer. The A-Lamp approach reaches a classification accuracy of 82.5% on AVA, showing that a saliency-driven choice of patches can bring substantial advantages over a random or fixed patch selection strategy. Sheng et al. (2018b) propose a multi-patch (**MP**) network with an attention mechanism (Stollenga et al. 2014): instead of using a pre-trained saliency model as in A-Lamp, the selection of salient patches in MP is learned directly from aesthetic labels, by assigning different weights to different image patches. Among the different weight assignment schemes considered, an adaptive one (MP_{ada}) obtains 83.03% classification accuracy on AVA. The state-of-the-art aesthetic classification methods in 2020 employ a combination of multi-patch networks, attention mechanisms and global features (Liu et al. 2020; Xu et al. 2020), achieving a classification accuracy of 83.59% on the standard AVA test set.

4.4.2 Content-Adaptive CNNs

As discussed in Sect. 4.2.3, considering the semantic content of a picture can help in assessing aesthetics. Compared to hand-crafted approaches, deep-learning-based methods can capture semantic information much better, and indeed many CNN architectures for aesthetic prediction employ the availability of additional content labels whenever possible (e.g., AVA provides additional information related to content and style, see Sect. 3).

The common way used in the literature to employ semantic information is to add a scene classifier in the deep model. A typical categorization used in aesthetics is based on 7 classes: *human, plant, architecture, landscape, static, animal* and *night*. These categories were initially proposed by Tang et al. (2013) and have been later used in many deep aesthetic models. The MNA-CNN network (Mai et al. 2016) discussed above includes a scene-categorization CNN fine-tuned on these 7 categories. Wang et al. (2016) build a multi-scene deep learning model (**MSDLM**) by cascading four convolutional layers of AlexNet (Krizhevsky et al. 2012), which is supposed to recognize the kind of scene, with a scene convolutional layer composed of 7 parallel convolutional blocks corresponding to 7 possible scene categories. The

scene group layers are pre-trained on images of a specific category to improve the classification performance. This work achieves an accuracy of 76.95% on AVA.

Another way to leverage semantic information of the scene consists of *multi-task* learning, in which a main task (aesthetics) is learned together with other additional tasks—in this case, a predictor of the image category (see Fig. 5b). Since both tasks are optimized concurrently in the network, the relative importance of the two task losses is a critical factor for a successful multi-task learning. Kao et al. (2017b) propose two possible solutions to determine the task weights. In their basic multi-task CNN architecture (**MT-CNN**), the relative importance of the aesthetic and semantic tasks is fixed to be $2/M$, where M is the number of categories ($M = 29$ semantic tags from AVA is used here). This network achieves an accuracy of 78.56% on AVA. The relative weights of the tasks can also be discovered directly from data, based on a Bayesian interpretation of multi-task learning. In particular, the relationship between tasks is embedded in the loss function under the form of a covariance matrix between the task-specific network parameters (corresponding to layers where parameters are not shared between tasks). The training procedure then consists of alternating steps of gradient descent and covariance matrix update. This network is called multi-task relationship learning CNN (**MTRL-CNN**). The classification accuracy with learned task weights rises to 79.08%. Despite the elegant mathematical formulation behind MTRL-CNN, the simultaneous calibration of the tasks remains challenging in practice, and later work has shown that training the network in two stages (by fine-tuning a semantic predictor) can lead to better aesthetics classification (Murray & Gordo 2017).

4.4.3 Aesthetic Regression

Providing a two-class aesthetic prediction may be insufficient in many applications where a finer-granularity assessment is desirable (e.g., for image enhancement). In those cases, it is more appropriate to estimate an *aesthetic rating* through a regression network. In particular, existing methods have focused on predicting the *average* score for an image, as given by human raters, e.g., a value between 1 and 10 for the AVA dataset. It is relatively straightforward to modify the architectures presented above to predict a continuous aesthetic score rather than a binary value. For instance, in Kao et al. (2015) the last layer of the network, which is a two-way softmax in aesthetic classification, is replaced by a single neuron to produce a scalar value. The loss used is the mean squared error. The performance criteria in the case of regression is no longer the accuracy, but rather measures such as mean squared error (MSE), root-mean-square error (RMSE), mean residual sum of squared errors (MRSSE), Pearson or Spearman rank-order correlation coefficients (PCC or SROCC, respectively). Nevertheless, it is typical to also provide classification results by thresholding the predicted scores, e.g., to the cut value of $5 \pm \delta$ in AVA (see Sect. 3.3), to benchmark the proposed methods with the state of the art. Current deep-learning-based methods for predicting the aesthetic mean score reach a correlation with ground-truth slightly in excess of 0.7, which is significantly

lower than the performance of no-reference technical quality assessment metrics, where the correlations are generally well higher than 0.8. This fact confirms the challenging nature of aesthetic quality assessment, but also raises some questions regarding the subjectivity of ground-truth scores (see Sect. 5.1).

To partially take into account the intrinsic subjectivity of aesthetics, a particular class of aesthetic regression networks aims at predicting the *distribution* of scores, rather than their mean. These systems include the popular neural image assessment (**NIMA**, Talebi and Milanfar (2018)), the aesthetic prediction model (**APM**, Murray and Gordo (2017)) and others (Jin et al. 2016a, 2018). We will discuss these techniques in more detail in Sect. 5.1.

4.4.4 Fusing Hand-Crafted and Deep Features

As discussed at the beginning of this section, an advantage of hand-crafted features over deep-learning-based methods is the intrepretability of aesthetic predictions. Some computational aesthetics approaches try to integrate the benefits of pure deep models and hand-crafted attributes by proposing mixed solutions fusing expert knowledge with data-driven features.

For example, Kucer et al. (2018) consider a mix of 331 hand-crafted features, obtained by some of the methods discussed in Sect. 4.2, and of deep features extracted by deep CNN such as VGG or ResNet. Using a tree-based learner, the authors show that, even if individually these feature sets are dominated in performance by current neural networks solutions, the (early or late) fusion of the features can provide competitive performance. In addition, the use of the tree-based learning approach allows one to deduce the importance of each feature in the aesthetic decision, and to significantly reduce the size of the feature set to less than 15% of the original size. The accuracy of this method on AVA is 81.95%, which is competitive with respect to more recent methods based on deep learning only. Notice that the explainability, i.e., which attributes are more relevant to the aesthetic decision, is achieved only in an average sense here, but not per picture.

A very different approach is that of Wang et al. (2017), who propose a deep network based on the Chatterjee's visual neuroscience model (Deep Chatterjee's machine, **DCM**) (Chatterjee 2003). The Chatterjee's model provides some insights on how humans perceive aesthetic quality: the human brain works as a multi-level system, in which the visual sensory input first processes a number of relevant features through a set of parallel pathways. Afterwards, the output of these pathways are associated and synthesized at a higher level into an aesthetic decision. Inspired by this framework, DCM computes several aesthetic attributes in parallel, using either hand-crafted features (in this case, simply the hue, saturation and value color representation), or CNNs which are trained in a supervised manner to predict one of the 14 AVA style labels (*complementary colors, duotones, vanishing point,* etc.). In a second step, a high-level synthesis network is used to fuse the attributes, and the overall network is trained to learn the distribution of votes (using the Kullback-Leibler divergence as metric). The authors also provide an interesting study on

the sensitivity of aesthetic prediction on the transformation of the input image (e.g., reflection, rotation, noise, etc.), which provides useful hints to perform data augmentation for aesthetics. The reported classification accuracy on AVA is 78.08%.

4.4.5 Learning an Aesthetic Ranking

The works that we have reviewed so far cast aesthetic prediction as either a classification or regression problem. In practice, often an aesthetic decision involves the comparison of two or more pictures, e.g., to decide which photo to keep in a personal album. It is clear that aesthetic classification is not sufficient in this case, and even a continuous rating might be imprecise when assessing the preference between two images. As an alternative, some works propose learning a ranking relationship directly from data, using a *ranking loss*.

Kong et al. (2016), who also proposed the AADB dataset (see Sect. 3), employ a Siamese network (Chopra et al. 2005) that takes as input a pair of images and directly predicts their relative ranking and aesthetic scores (see Fig. 5c). The network is constituted by two identical branches with shared weights, and is trained by minimizing the following *contrastive* loss term:

$$\mathcal{L}_{\text{contrast}} = \sum_{i,j} \max\left(0, \alpha - \eta(y_i \geq y_j)(\hat{y}_i - \hat{y}_j)\right), \tag{2}$$

where y_i and $\hat{y}_i$ are the ground-truth and predicted average rating for image i, $\eta(y_i \geq y_j) = 1$ if $y_i \geq y_j$ and $\eta(y_i \geq y_j) = -1$ otherwise, and α is a margin parameter. The contrastive loss penalizes predictions that invert the original aesthetic ranking of images more than predictions that preserve this ranking. In this second case, predictions that provide the correct ranking and estimate scores spaced out by at least the margin α are less penalized to focus the learning process on the difficult pairs with similar ratings. In addition to the contrastive loss, a regression term (e.g., MSE) is also added to anchor the predicted scores to the original rating scale. This basic Siamese architecture is integrated into an attribute and content-adaptive network, and experiments show an overall SROCC of approximately 0.56, and a classification accuracy of 77.33% on AVA. Performance on AADB is higher (correlation in excess of 0.67). Interestingly, the authors also provide a cross-dataset train/test evaluation, showing that a network trained on AADB has very poor performance (SROCC ≈ 0.15) on AVA, and vice-versa. This opens up a number of questions regarding the generalization capabilities of deep-learning-based aesthetic predictors.

A different ranking loss is employed in Schwarz et al. (2018), which uses a *triplet network* architecture to learn an aesthetic distance in the feature space (Hoffer & Ailon 2015). Compared to the Siamese architecture, the triplet network has three columns with shared weights, which receive three inputs: an anchor image a, an

aesthetically similar image p and an aesthetically dissimilar image n. The network is trained by minimizing a triplet loss:

$$\mathcal{L}_{\text{triplet}} = \sum_{a,p,n} \max\left(0, \alpha + \|\Phi_a - \Phi_p\|_2^2 - \|\Phi_a - \Phi_n\|_2^2\right), \tag{3}$$

where Φ_a, Φ_p and Φ_n are embeddings (i.e., deep CNN features in this case) for a, p and n, respectively, and α is a margin parameter. Intuitively, the triplet loss pushes images that have similar aesthetic level to be close in the feature space, and images which have very different aesthetic values to have very different embeddings, thus enforcing a ranking among images. The reported results of the fine-tuned network on AVA do not show significant improvement over the Siamese architecture described above (accuracy of 75.83%), although the two networks are not comparable as Schwarz et al. (2018) does not include attribute and semantic information.

To conclude this section, we report in Fig. 6 the classification accuracy on the AVA dataset of some of the deep-learning-based methods discussed above. We can clearly see a performance improvement (over 10% gain) in accuracy in the past six years. Also, we observe that performance have been saturating in the last years to slightly less than 84% when $\delta = 0$ is used to label the aesthetic classes in AVA. It seems difficult nowadays to go far beyond this value using the AVA dataset. This limit raises questions regarding the nature of aesthetic data used as ground-

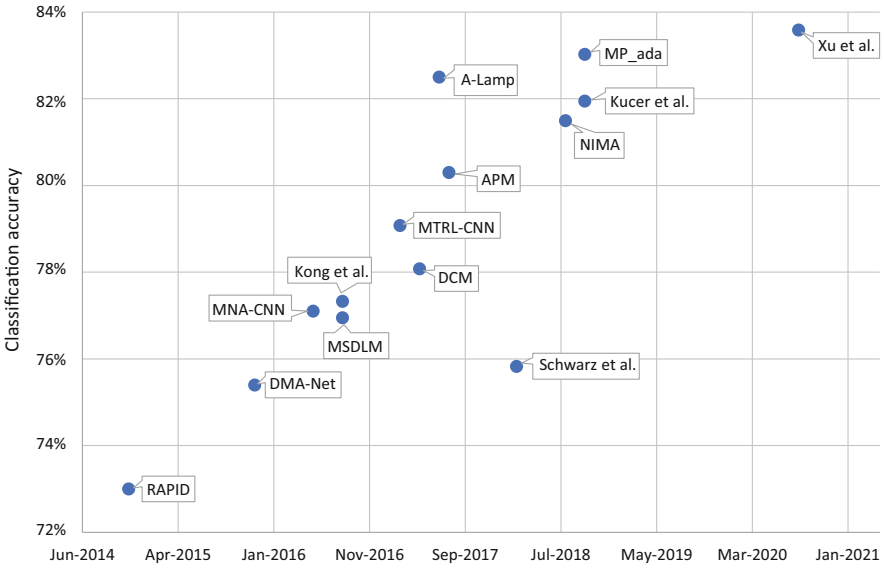


Fig. 6 Evolution of binary classification accuracy on the AVA dataset for some relevant deep-learning-based methods

truth: as discussed in Sect. 3, the aesthetic scores crawled from DPChallenge can be significantly influenced by the semantic context (challenge, content, etc.), which makes the ground-truth scores irremediably noisy and affected by other factors than aesthetics, such as interestingness. How to collect large aesthetic datasets with clean labels is still an open question, and only little work has been devoted to it in the multimedia community, compared to the more traditional technical quality assessment problem, for which guidelines and recommendations have been available for several decades (e.g., ITU-R (2012)).

5 Challenges in Computational Aesthetics: Subjectivity and Explainability

The overview of computational aesthetic methods presented in the previous section demonstrates that substantial progress has been made in this field in the last 15 years. However, it also points out some limitations and weaknesses of the current state of the art in computational aesthetics. In addition to the still limited accuracy of aesthetic prediction approaches, we have already mentioned some open challenges in the field of computational aesthetics, including the reliability of the ground-truth scores, the capability to explain the aesthetic judgments, and the subjective nature of aesthetic decisions. In this section we discuss these challenges, and in particular the dimensions of *subjectivity* and *explainability* in computational aesthetics.

5.1 Dealing with Subjectivity

In Sect. 1 we have introduced the classical subjectivist/objectivist debate in aesthetics. As we have mentioned there, the vast majority of existing computational aesthetics methods embrace an objectivist hypothesis on the aesthetic quality of photos. Specifically, they assume beauty is a property of the picture, produced by a combination of its attributes, which is essentially belonging to the object rather than the observer, thus being *universal*. This hypothesis legitimates the identification of an aesthetic score as a pooling operation over a set of opinions (e.g., average, or majority vote, etc.), which is taken as the ground truth of aesthetic prediction.

In practice, while opinions of multiple observers might follow a common trend, individual opinions are inherently subjective. The causes of this *subjectivity* are varied. They can be imputed to the inner state of the viewer and his/her contingent feelings, mood, sensations, etc. In photography, subjectivity can occur due to different evaluation criteria followed by photographers (Barrett 2020), which are also influenced by the historical epoch, cultural context and demographics of the observer (Kairanbay et al. 2019; Redi et al. 2016). The level of expertise of the viewers can also impact the perception of aesthetics (Lebreton et al. 2016), e.g.,

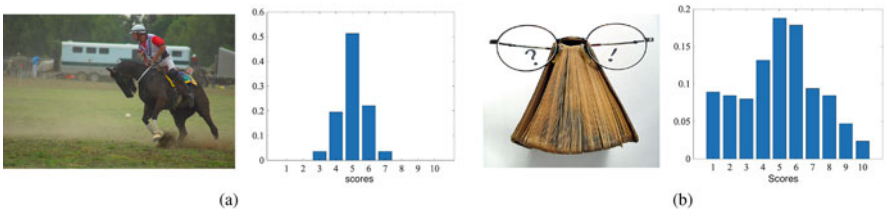


Fig. 7 Subjectivity in image aesthetics. The two photos (taken from the AVA dataset) have exactly the same average aesthetic score. However, their normalized score distribution (displayed on the right panels) reveals a very different degree of consensus of human raters. (**a**) A photo with low subjectivity. (**b**) A photo with high subjectivity

functional magnetic resonance imaging (fMRI) scans have revealed significant differences in the neural activities between architects and non-architects when evaluating photos of buildings (Kirk et al. 2009). A study carried out using magnetoencephalography has discovered significant differences in brain activity when assessing the beauty of photos and paintings in male and female participants (Cela-Conde et al. 2009).

Due to subjectivity, the opinions of individual viewers may be in disagreement with each other. We define the aesthetic subjectivity of a picture as the *degree of consensus* about its aesthetic value when this is judged by a panel of human observers (Kang et al. 2019). Figure 7 illustrates this definition with two example images from the AVA dataset. Compared to the traditional technical quality assessment, where inter-viewer agreement is generally high, in aesthetics the human judgments tend to be more dispersed. In the following, we discuss some attempts to include the subjectivity dimension in computational aesthetics.

5.1.1 Predicting Score Distributions

A popular way to consider aesthetic subjectivity is to predict the *distribution* of the image aesthetic scores. This is represented as a vector of probabilities over a set of ordinal values instead of a single one-dimensional estimate (e.g., average score or the aesthetic class). Predicting score distributions requires adapting computational aesthetics techniques to process categorical probability distributions as labels. In particular, while conventional loss functions may be used (e.g., the Huber loss is used to reduce the impact of outliers by Murray and Gordo in the APM network (Murray & Gordo 2017)), algorithms to predict score distributions employ different loss terms for training. More specifically, employing a simple vector distance such as the L2 norm between histogram vectors is in general sub-optimal, as it does not consider the ordinal nature of the aesthetic ratings. For example, given a reference score distribution on a 5-level discrete scale $p_1 = (1, 0, 0, 0, 0)$, the two following score distributions $p_2 = (0, 1, 0, 0, 0)$ and $p_3 = (0, 0, 0, 0, 1)$ have the

same Euclidean distance from p_1. However, it is intuitive that p_2 is closer to p_1 than p_3, since the aesthetic scores where the probability mass is concentrated are closer.

Among distances between probability distributions, one that has been widely used in aesthetics is the *Earth mover's distance* (EMD). For two discrete distributions p and q, the EMD is computed as the L2 norm of the difference between their corresponding cumulative distribution functions (cdf) P and Q, that is:

$$\mathrm{EMD}(p,q) = \left[\sum_{i=1}^{K} P(i) - Q(i)\right]^{\frac{1}{2}}, \tag{4}$$

where K is the number of score levels (e.g., $K = 10$ for AVA). By employing the cumulative distributions, the EMD is sensitive to the order of the probability masses. The use of EMD to predict aesthetic score distributions was first proposed by Wu, Hu and Gao in 2011 (Wu et al. 2011). They introduce a modified support vector regression algorithm called support vector distribution regression (SVDR), trained with a squared EMD. In addition, they also proposed a weighting mechanism to penalize more errors on images which have a reliable ground-truth score distribution, called reliability-sensitive learning (RSL). The reliability is measured as the number of votes received by the image: the higher the number of votes, the closer the sample histogram is to the true population distribution. The EMD has been later used by other aesthetic prediction methods, including the popular NIMA system (Talebi & Milanfar 2018). Similar ideas to Wu et al. (2011), in particular the reliability term, have been employed by others afterwards, e.g., it has been integrated in a label distribution learning framework in Cui et al. (2017) (however, a hinge loss is used there).

Other distances between probability distributions can be considered. For instance, Jin et al. (2016a) predict aesthetic histograms via a modified VGG-16 network trained with the χ^2 (Chi-square) distance, defined as:

$$\chi^2(p,q) = \frac{1}{2}\sum_{i=1}^{K} \frac{(p_i - q_i)^2}{p_i + q_i}. \tag{5}$$

This distance gives less importance to the difference between large bins, and was successfully used for texture and object classification, local descriptor matching, etc. (Pele & Werman 2010).

Another family of methods to predict aesthetic distributions employs distances (or, more precisely, pseudo-distances) borrowed from information theory. We already mentioned the Deep Chatterjee's Machine (DCM, Wang et al. (2017)) in Sect. 4.4.4. It approximates the underlying aesthetic distributions as Gaussians, and measures their distance with the *Kullback-Leibler* (KL) divergence, which in this case has a simple closed-form expression:

$$\mathrm{KL}(p,q) = \log\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\mu_q^2} - \frac{1}{2}, \tag{6}$$

where $\mu_p, \mu_q, \sigma_p, \sigma_q$ are the means and standard deviations of p and q, respectively. The Gaussian approximation of p and q does not just allow simplification the computation of the KL divergence, but also solves the issue of defining the KL for values with zero probability mass. However, the hypothesis of normality of the distributions seems somewhat too strong, at least for the AVA dataset: although in Murray et al. (2012) it is found that most images in the dataset have an approximately Gaussian distribution of scores, later studies (Park & Zhang 2015) have shown that the distributions are better approximated as power laws. Indeed, this is even more evident for the images with extreme scores, which have skewed distributions. Another drawback of Eq. (6) is that the KL divergence is *asymmetric* ($\mathrm{KL}(p,q) \neq \mathrm{KL}(q,p)$). To overcome this limitation, the KL divergence is often symmetrized as $\mathrm{KL}_{sym} = \frac{1}{2}(\mathrm{KL}(p,q) + \mathrm{KL}(q,p))$.

To consider the ordinal nature of the ratings and solve the asymmetry of the KL divergence, Jin et al. (2018) employ a cumulative Jensen-Shannon divergence (CJS) loss. The Jensen-Shannon divergence is a symmetrized KL divergence of the two distributions p and q with respect to their average $m = \frac{1}{2}(p+q)$. In CJS, the Jensen-Shannon divergence is computed on the cumulative distributions P and Q:

$$\mathrm{CJS}(p,q) = \frac{1}{2}\left[\sum_{i=1}^{K} P \log \frac{P(i)}{M(i)} + \sum_{i=1}^{K} Q \log \frac{Q(i)}{M(i)}\right], \tag{7}$$

where M is the cdf of m. In addition to the plain CJS loss, the authors also include a reliability weight inspired by Wu et al. (2011), with the difference that they use the *kurtosis* of the ratings distribution instead of the number of voters. The use of kurtosis as a measure of subjectivity was proposed also in Park and Zhang (2015) before (see next section).

Although predicting score distributions can provide complete information about aesthetic consensus, the predicted distributions are in practice difficult to interpret. Since evaluating the prediction of histograms requires choosing a distance metric between distributions, comparing the results of different methods may not be conclusive. In fact, to validate the proposed approach, these works often resort to extracting simpler aesthetic measures such as the average or aesthetic class from the estimated distributions, in order to compare to the state of the art. In addition, the ratings in large aesthetic datasets such as AVA tend to concentrate around the middle quality (see Fig. 2a). As a result, most of the training samples have a distribution that is approximately Gaussian and centered around the middle score. This over-representation of images with mediocre quality leads to a sort of "center bias" in the prediction: the estimated distributions tend to resemble the average score distribution of the dataset, entailing poor prediction performance for images with very high or low quality. This phenomenon occurs as well for mean score regression, and a traditional solution in aesthetics consists of excluding images with average ratings close to the middle of the rating scale from training (Datta et al. 2006; Ke et al. 2006; Lu et al. 2014). Another option to mitigate the score imbalance consists of using resampling or a weighting scheme to balance the loss during training. For

example, in Jin et al. (2016a) the weights are computed as the inverse of the (binned) distribution of the average aesthetic score over the AVA dataset. In this way, less frequent scores are assigner larger weights and are penalized more during training, thus effectively driving the network to focus on rare samples.

5.1.2 Measures of Subjectivity

While predicting the distribution of aesthetic ratings gives an idea of the consensus of human observers on the quality of a picture, in many cases it is desirable to extract a single, scalar measure of subjectivity, e.g., to be used as a quality metric or a penalty term in an optimization or learning process. A few works have addressed this problem, by computing some significant statistic based on the rating distributions (e.g., the variance or higher-order moments), and evaluating its prediction through machine learning approaches.

Kim et al. (2020) study the objectivity and subjectivity in aesthetic quality assessment. The "objectivity" is identified with the task of predicting the mean aesthetic score or an aesthetic class, which corresponds to the classical setup in computational aesthetics and to the perspective we have taken in the previous part of this chapter. The term subjectivity, instead, is quantified as the *standard deviation* (std) of the scores. Based on these definitions, the authors propose a prediction scheme for the two terms. They first crawl a new database from DPChallenge containing more than 300k pictures posted over a time interval of 12 years. This long time horizon allows the authors to make some interesting observations regarding the evolution of objectivity and subjectivity: e.g., due to the increase of the photographic device quality, the average aesthetic scores in DPChallenge have increased with time, while the average subjectivity has decreased. Afterwards, the authors extract 295 features from each image, which are combined through an SVM to predict either the mean or the std of the scores. Through a feature selection process, it is also possible to understand which are the most significant features in each of the two tasks. Notice that both objectivity and subjectivity here are *quantized to two classes*, i.e., the prediction is a binary classification problem. The separation into two classes discards images with medium std values (similar to what is typically done on mean scores with the parameter δ). Under these assumptions, the classification accuracy for the mean score prediction is 71.6%. For std, it lowers down to around 67%, with larger inter-category variations (e.g., for landscape images the std prediction accuracy exceeds 77%, while for architecture it is around 61%). While overall std prediction seems more difficult, the results are encouraging, showing that predicting subjectivity is feasible. The authors also investigate the sources of subjectivity through an analysis of text comments associated to the images (downloaded from DPChallenge). The "unusualness" and the coexistence of both aesthetic merits and defects explains the high levels of subjectivity.

The conclusion that subjectivity can be predicted with reasonable accuracy is somehow contradicted by the work of Kang et al. (2019), although the results cannot be directly compared as the evaluation schemes are different (regression in this

case). The correlation coefficient between the predicted std and the ground-truth is only ≈ 0.3, compared to correlations in excess of 0.7 obtained by state-of-the-art methods to predict the mean aesthetic score. We hypothesize that the higher performance in Kim et al. (2020) is significantly influenced by the removal of samples with medium std values, which are the most significant portion of the data (see Fig. 2b). The authors of Kang et al. (2019) also propose other subjectivity measures in addition to std, including two novel measures based on information theory. These measures compute the distance of the ratings distribution of an image to an ideal distribution having maximum entropy (and thus, minimum consensus). Even if these new measures can be predicted slightly better than std on the AVA dataset (which may imply they are more robust to noise), the overall prediction performance remains poor, most probably due to the complex, contextual factors leading to little aesthetic consensus.

Park and Zhang (2015) present an original and very interesting analysis of the consensus in aesthetics (in particular, for the AVA dataset). Instead of using the variance of the scores, which is seriously distorted by highly skewed and bounded data, they consider the fourth moment of the distribution, i.e., *kurtosis*, as an indicator of subjectivity. Kurtosis measures how long are the tails of a distribution. The kurtosis of a distribution is linked to its skewness by the relation: $\text{kurtosis} \geq (\text{skewness})^2 + 1$. Therefore, to characterize subjectivity, Park and Zhang study the distributions of images in the skewness-kurtosis plane—a representation they call *SK maps* (see Fig. 2c), which has been used in physics and finance to study the deviations from Gaussianity. The SK maps provide insightful information about the subjectivity of images in AVA. First, it is observed that there is a strong non-Gaussianity in the scores of the AVA images. In particular, images with average scores around 5 tend to have a wide range of kurtosis, which implies they follow very different (and non Gaussian) distributions. In addition, images with low aesthetic scores (i.e., with positive skewness) tend to have higher kurtosis, i.e., there is more aesthetic consensus in judging aesthetically unpleasing pictures than high-quality ones. Finally, the SK maps differ significantly based on the content category, which is coherent with the content-dependent subjectivity observed in other works afterwards.

Based on the SK map representation, Park and Zhang also present a mathematical *dynamic model* to explain subjectivity in aesthetic perception. The approach is based on the classical drift-diffusion model, previously used by psychologists to explain behavioral data in emotion analysis tasks. The drift-diffusion model assumes that, in the absence of any external stimulus, the human mind performs an internal random walk. When a decision between two or more options is to be made, the brain accumulates evidence favoring each of the alternatives over time. The combination of these "clues" (attractors) with the noise component (random walk) can be depicted as a particle drifting and diffusing between two boundaries, until it reaches one of them. Similarly, when the aesthetic judgment converges to one state (e.g., good or bad aesthetic quality), the aesthetic decision is taken. This simple drift-diffusion model allows the explanation of most of the behaviors observed in the SK maps, and provides a foundation for results obtained by later studies (Kim

et al. 2020). In particular, when multiple, balanced attractors are present (i.e., both positive and negative aesthetic attributes), the judgment tends to converge towards a mediocre aesthetic score. Moreover, the convergence time is longer, i.e., humans employ a longer time to evaluate images with larger subjectivity. This conclusion is supported by a user study in which the authors recorded the voting time. Even more interestingly, the drift-diffusion model suggests that it is the mixture of positive and negative attractors in a training sample that misguide most machine learning methods, making the subjectivity prediction performance poor. Instead, since subjectivity is the result of a dynamic system, a proper learning scheme should embed this dynamic aspect, e.g., using an active learning approach. Unfortunately, this original perspective, which might open new directions in the understanding of aesthetic subjectivity, has not been further investigated in follow-up work on computational aesthetics.

5.1.3 Personalized Aesthetics

A different approach to subjectivity in computational aesthetics departs substantially from the methods that we have analyzed so far in this chapter. Instead of focusing on the *universal* scope of aesthetics (see Fig. 1), we briefly describe in the following some methods that aim at predicted *personalized* aesthetics for a particular person. As we mentioned in Sect. 1, personalized computational aesthetics assumes an interactionist interpretation of aesthetics, where the individual perception is the result of the interaction between some objective, intrinsic features of a photo, with a subjective processing/interpretation.

Personalized aesthetics algorithms aim to adapt a generic aesthetic predictor to the individual tastes of a person, based on the availability of a small set of annotations from that user. To this end, they employ tools often used in image recommendation and user profiling, such as active learning, collaborative filtering or residual learning. Park et al. (2017) propose a joint regression and ranking algorithm to score and rank a set of user-specific images $\mathcal{T}$. The system first extracts a subset $\mathcal{S}$ of training images from a general aesthetics dataset (e.g., AVA). The images to extract are selected as the nearest neighbors to the images in $\mathcal{T}$. In a second phase, the user ranks a small subset of images in $\mathcal{T}$. Finally, combining these two sources of information, the system learns to predict all the scores and ranks in the remaining images of $\mathcal{T}$. The authors use a max-margin learning algorithm, in particular, an SVR (inputted with a feature vector of 4096 elements, extracted from the second last layer of AlexNet (Krizhevsky et al. 2012)) for learning the universal aesthetic part, and a ranking SVM (R-SVM) to learn a ranking model given the partial orders on the training data. The two losses are combined to jointly learn a *ranking support vector regression* (R-SVR). The results, validated by a user study, are promising and show that the proposed approach can produce cleaner ranking predictions compared to a general aesthetic model alone.

Ren et al. (2017) make similar hypotheses, in particular, that only a small number of annotated examples from a user is available. To be able to still learn significant

personalized aesthetic scores in this setting, they adopt a residual-based model adaptation scheme to *learn a scalar offset* to the generic aesthetic score predicted by a universal aesthetic predictor. The authors start by collecting two datasets: one is FLICKR-AES, containing 40k Flickr images rated by 210 unique AMT annotators; the other is REAL-CUR (Real Album Curation Dataset) which contains 14 real users' photo albums with aesthetic scores provided by the album owners. Afterwards, they estimate aesthetic attributes (with a network fine-tuned on the AADB dataset attributes) and the image category (content class) for each image in FLICKR-AES. An analysis on these results and the ground-truth user preferences reveals strong correlations between personal preferences and attributes/content of an image. This observation is key for the proposed approach: in fact, predicting a score offset using an end-to-end optimization would be unfeasible, given the very small percentage of images annotated with individual preference. Instead, the predicted attributes and classes, represented as 10-dimensional categorical distributions (obtained by the last softmax layer in the attribute and content prediction networks) are used as input features for an SVR to predict an offset for a given image. This system is also extended to an active learning scenario, where the model is updated while the users evaluates new images; in this case, the choice of the images to score can be optimized according to heuristic criteria.

In some circumstances, collecting extra labels for specific users to perform personalization is impractical or time consuming. A simpler alternative consists in sensing user-specific aesthetic preferences from the user's personal favoring behavior on social media platforms. Cui et al. (2020) leverage this idea and collect personalized preferences from a set of 50k professional photos downloaded from Flickr. Photos are considered "professional" if they have been posted by one of the top 200 photographers in the ranking of the website. Analyses on this image set show that users tend to prefer images which have some common aesthetic features. However, learning personal preference on this dataset is difficult as, on average, users favor only a very small portion of the total number of images. Therefore, similar to the works discussed above, the authors learn first a universal aesthetic model to extract meaningful aesthetic features. Afterwards, they use a *collaborative filtering* approach to minimize a twofold objective: on one side, a pairwise loss term to guarantee that the user-specific ranking on favored vs. non-favored is respected (under the hypothesis that a favored picture is aesthetically better for the user); on the other hand, a regularization term to smooth out the predicted scores in such a way that they are not too distant from the average ratings. As the authors also point out, the major pitfall of this approach is in the assumption that "faves" approximate somehow the aesthetic value of a picture. Nevertheless, as we have discussed throughout this chapter, this assumption is often made in computational aesthetics to collect data at low cost, even though it can lead to noisy prediction and hardly interpretable results.

5.2 *Explaining Aesthetic Scores*

While the mainstream aesthetic research has focused on improving the prediction of aesthetic scores or classes, relatively little has been done to understand *why* an image is aesthetically pleasing or not. This question is particularly challenging for deep-learning-based methods, due to the very high dimensionality of the employed models that make them significantly hard to interpret. Nonetheless, some works have tried to analyze the predictions of neural networks in aesthetics, or to justify the aesthetic scores by producing explaining text comments. Moreover, some datasets have been collected with the specific purpose of providing extra ground-truth labels to facilitate aesthetic explainability.

5.2.1 Visualization Techniques

An approach to explain aesthetic scores obtained by a convolutional neural network consists in analyzing the filters and the features learned by the network. This category of methods has been quite popular in computer vision in the early stages of development of deep CNN to visualize what the network was learning (Zeiler & Fergus 2014). For instance, analyzing the filters at different layers of a classification network shows that initial layers perform low-level filtering (e.g., gradients, Gabor filters, etc.), while deeper layers are optimized to capture higher-level structures and parts of objects. This kind of visualization has been also applied to networks that predict aesthetics (Kao et al. 2016; Jin et al. 2016b). However, the conclusions from this inspection are in general very limited, as the learned patterns reflect the same kind of behavior observed in non-aesthetic networks, making them difficult to be interpreted.

Another technique to analyze the features learned by a CNN is to study *class activation maps* (CAM) (Zhou et al. 2016). In the simplest setting, CAMs can be obtained for classification networks satisfying a particular structure, i.e., having a global averaging pooling layer followed by a single fully connected layer before the output layer. In this case, for a given input image and a certain class, the score of the class is mapped back to the previous convolutional layer to generate a corresponding class activation map. CAMs can be visualized as low-resolution images, which highlight the class-specific discriminative regions. Later work (e.g., Grad-CAM (Selvaraju et al. 2017)) extends this visualization technique to a much wider variety of networks, by propagating back the gradient of a target class to a convolutional layer of the net. Class activation maps have been employed also in the case of computational aesthetics. Kairanbay et al. (2017) build on the CAM visualization to provide a justification of high vs. low aesthetic quality. They observe that aesthetically pleasing images tend to have activation maps with energy well concentrated around salient objects of the picture. Conversely, photos belonging to the low-quality class have activations that are spread around the picture and on non-interesting regions. The authors speculate that this behavior reflects basic rules of

photography, such as the importance of focusing on the subject and the concept of negative space (see Fig. 3). However, such observations are verified qualitatively only on a few images, and it seems difficult to generalize this conclusion to more complex scenes or photos where the subject is not clearly identified. Zhang et al. (2018) extends this analysis by visualizing activations at different levels of a multi-task network that predicts simultaneously an aesthetic class and one of the 66 AVA semantic tags. Thus, in addition to activation maps for aesthetic attributes, they also study CAMs for attributes. Jointly predicting the activation maps for the two tasks has the potential to not only localize aesthetically salient areas in the picture, but also to explain why they are important (by intersecting the two maps). However, the conclusions remain still vague and difficult to justify when considering a wide variety of content. An interesting application of computing activation maps for aesthetics consists of automatically cropping a picture by keeping the most aesthetically relevant regions (Kao et al. 2017a; Zhang et al. 2018).

Murray and Gordo (2017), whose APM model we have introduced earlier, employ a different visualization technique compared to CAM. They leverage the concept of *adversarial examples* (Goodfellow et al. 2014), i.e., input samples that are imperceptibly modified to completely alter the prediction of a network, while looking essentially the same to a human observer. Based on this concept, they change the score distributions of test images to be slightly better or worse than the original sample. Then, they modify the image by gradient descent in such a way to obtain a new image that matches the altered distribution. Visualizing which pixels have been modified in the original test image in order to improve or reduce aesthetic scores provides an indication of the regions of the picture that are used by the model to make predictions. Compared to CAM representations, this technique allows one to obtain higher-resolution visualizations. The authors notice that most changes are localized in salient regions of the pictures, confirming observations from previous work. However, an inspection of the error images leaves still many open questions about the interpretability of these maps. In addition, the adversarial examples demonstrate that even imperceptible modifications in the original pixels can yield significant changes in the image scores. This fact indicates that aesthetic networks are also prone to adversarial attacks as other computer vision applications such as object classification, and raises some fundamental questions about how much neural-network-based computational aesthetic predictors are reliable.

5.2.2 Generating Text Explanations

As we have discussed above, aesthetic explanation approaches based on network activation maps or other visualization techniques alone have not been able so far to provide convincing evidence of why a given picture is beautiful or not. A more explicit approach to generate plausible explanations consists of producing a text comment about the qualities and defects of a photo.

We have already discussed in Sect. 4 a few seminal works linking aesthetic quality not only to pixel-based characteristics, but also on associated textual comments

from users (San Pedro et al. 2012; Marchesotti 2013). The considerable progress that deep learning techniques have brought to natural language processing (NLP) has enabled the use of advanced image captioning techniques in computational aesthetics. One of the first works in this direction is the one of Chang et al. (2017). They propose a *multi-aspect* aesthetic captioning system, where more than one aspect of an image can be commented, e.g., composition, color arrangement or subject contrast. This approach has a very reasonable foundation: in fact, it mimics some earlier studies in computational aesthetics that tried to decompose the global quality as a combination of some basic attributes (Aydın et al. 2014). The authors propose two architectures, both based on CNN-LSTM (long short term memory units) to produce a set of captions for a given image. It has to be noted as well that the authors also offer a new dataset with aesthetic captions crawled from a professional photographers website (https://gurushots.com/), called the photo critique captioning dataset (PCCD), see Table 1.

Wang et al. (2019) combine aesthetic classification and captioning into a multi-task network called *neural aesthetic image reviewer* (NAIR). This work leverages a dataset of 40k images extracted from AVA (AVA-reviews, see Table 1), that the authors collect based on images with text comments in AVA. To select images, they remove aesthetically ambiguous pictures ($\delta = 0.5$). The proposed network includes a part for image aesthetic classification based on a single-column CNN, and a part for vision-to-language generation that generates natural-language comments using a sequence of LSTM units.

Recently, Ghosal et al. (2019) have proposed a new dataset with 230k images and 1.5M captions for aesthetic image captioning called AVA-Captions. The dataset is obtained by cleaning the raw comments in AVA to retain the most discriminative n-grams, which are then used to train a CNN-LSTM network in a weakly-supervised way. The labels for training are obtained by processing the filtered captions, in such a way to extract terms corresponding to different attributes. However, instead of using fixed attributes as in Chang et al. (2017), here the attributes are discovered from data using *latent Dirichlet allocation* (LDA), a generative probabilistic model used in text modeling and retrieval. LDA clusters semantically similar terms, which correspond to classes of images (e.g., faces, landscapes, etc.). The discovered attributes go beyond the typical aesthetic attributes (color, contrast, composition, etc.) and include some semantic labels (e.g., “sky”, “sport”,“action shot”), but also opinions and judgments on the content (e.g., “cute expression”, “great action”). The generated captions display more diversity than those obtained on the noisy (original) captions from AVA, which tend to be monotonous and repetitive. The captioning is evaluated through a subjective experiment, showing a relatively good agreement with human opinions about the quality of a caption, which is mainly intended here as the informativeness and naturalness of the generated comment. Unfortunately, the produced text explanations depend significantly on the quality of the original captions, and judging their aesthetic relevance remains still an open problem.

5.2.3 Datasets with Aesthetic Attributes

The techniques to explain aesthetics based on data visualization or captioning described above can provide hints on the relevant regions or aspects of a photo. However, several drawbacks are related to these methods, particularly the difficulty of assessing their performance and their significant dependence on the input training data (especially for generated comments). These observations bring us back to a fundamental challenge in computational aesthetics, which we have mentioned many times throughout this chapter: collecting large-scale datasets with reliable, clean, and rich labels. At the time of this writing, there is still no aesthetic dataset able to provide, at the same time, a large number of annotated images *and* reliable, high-quality aesthetic scores. We have already discussed the features and limitations of some popular aesthetic datasets in Sect. 3. To study aesthetic explainability, aesthetic datasets should be complemented with additional information, e.g., aesthetic attributes to explain why an image is aesthetically pleasing or not.

Few datasets in the literature have explicitly elicited aesthetic attributes information from human raters. A notable example is AADB (Kong et al. 2016), where images are annotated with 11 aesthetic attributes. These attributes were defined based on expert knowledge: professional photographers were consulted to define a set of attributes that span the main dimensions of photography (color, light, composition, focus) and that can provide a natural vocabulary for practical applications from photo editing to retrieval. The set of selected attributes include: "interesting content", "object emphasis", "good lighting", "color harmony", "vivid color", "shallow depth of field", "motion blur", "rule of thirds", "balancing element", "repetition", and "symmetry". These attributes are assigned binary labels by each user. While the AADB attributes have an aesthetic valence, it is not clear whether they are sufficient to capture the wide range of factors that concur to form an aesthetic judgment. In addition, images in AADB are rated by only 5 users, which makes it difficult in practice to compute significant mean attribute values.

Recently, Kang et al. (2020) have proposed an Explainable Visual Aesthetics (EVA) dataset, which aims at partially solving the issues of AADB and other similar datasets. An example of the voting interface is illustrated in Fig. 8. In EVA, attributes are simplified to four general categories: "light and color", "composition and depth", "quality" (intended as technical quality), and "semantics". The attributes span different levels of factors affecting image aesthetics, from perceptual (light and color, technical quality), to photographic technique (composition, depth) and interpretation of the scene (semantics). Compared to AADB, the attributes are less detailed, and thus the information about why an image is beautiful is more generic. However, they are more inclusive and general, which might be beneficial to describe factors which are outside the vocabulary pre-defined by the experimenters. In addition, EVA attributes have two measurements: one to gauge the attribute magnitude (on a Likert scale) for a given image; the other to assess the attribute relevance (on a binary scale) in producing the overall aesthetic score. In addition to image aesthetic scores and attributes, EVA collects also the "difficulty" encountered by the user to rate an image, which is somehow related to the personal aesthetic

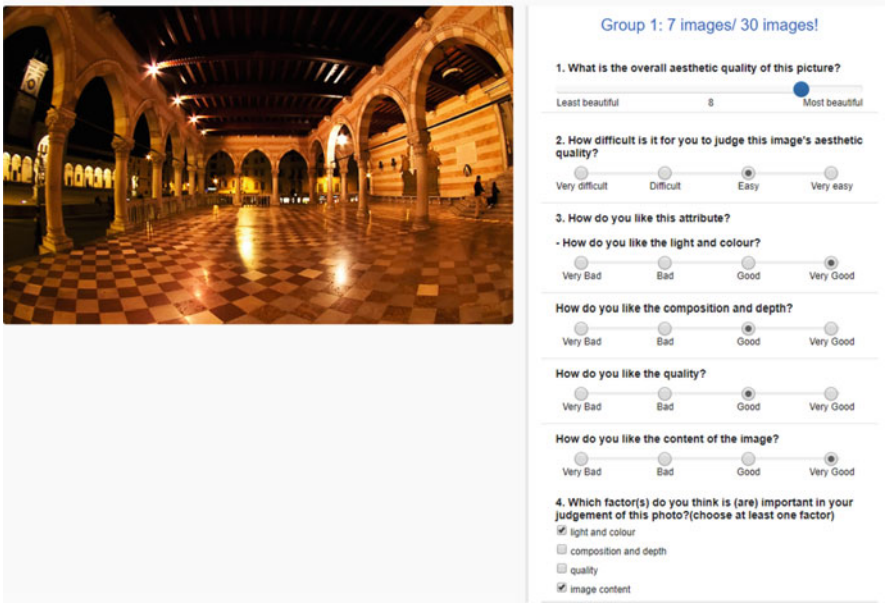


Fig. 8 Voting interface in the EVA dataset. In addition to aesthetic scores (discrete 11-levels scale), additional aesthetic attributes are collected (using 4-levels Likert scales), as well as their relevance (on a binary scale) to forming the overall aesthetic quality

uncertainty and might have interesting links to the study of subjectivity discussed earlier in this section. Furthermore, differently from previous datasets, the data collection in EVA includes a detailed training phase, in which raters are instructed about the meaning of attributes (with visual examples) and on how to use the rating scales, following common guidelines widely adopted in technical quality assessment (ITU-R 2012). EVA includes 4070 images, which is less than half of the images of AADB; however, each image has at least 30 votes. Despite the limited number of images, and the possible noise in the labels due to the crowdsourcing acquisition, the EVA dataset represents in our opinion a good starting point for further work on collecting better ground-truth labels for computational aesthetics.

6 Concluding Remarks

Computational aesthetics is a challenging and rapidly evolving field, at the intersection of multimedia quality, human perception and machine learning. In this chapter, we have given a general overview of this domain, from the philosophical debates around the interpretations of aesthetics, to the modern techniques to predict human aesthetic judgments. After the initial attempts to formulate aesthetics as

a mathematical object by Birkhoff in the 1930s, computational aesthetics has undergone an incredible development, in particular with the rise of data-driven methods in the past 15 years. We have discussed the fundamental role that datasets play in understanding aesthetic evaluation, and the different dimensions that should be taken into account when approaching computational aesthetics (focusing in particular on general aesthetics).

Computational methods to predict aesthetic classes based on deep neural networks can nowadays achieve a binary prediction accuracy higher than 83% on the benchmark AVA dataset (Murray et al. 2012). The classification performance on this dataset has now reached a plateau, in which it seems difficult to substantially improve predictions by just changing the architectures of the networks used. We have argued that this limit is somehow related to the noise in the aesthetic scores collected by crawling amateur or professional photography websites, as well as the intrinsic *uncertainty* of aesthetic evaluation, which is subjective in nature. We have thus pointed to some fundamental challenges in modern computational aesthetics: dealing with the subjectivity of the aesthetic scores; explaining aesthetic decisions; and building clean and reliabile large-scale datasets.

We conclude the chapter by mentioning that, in addition to the topics covered here, there are several other aspects related to aesthetics that could be further considered. In particular, in addition to numerous applications of image aesthetics to enhancement, recommendation, etc., mentioned throughout the chapter, we need to mention video aesthetics (Yeh et al. 2013; Bhattacharya et al. 2013) and related applications (e.g., thumbnailing (Song et al. 2016)), and finally recent studies linking brain-computer interfaces to the generation of aesthetically pleasing pictures (Spape et al. 2021), which appear to be a promising avenue to understand and predict aesthetic judgment mechanisms.

References

Amirshahi, S. A., Denzler, J., & Redies, C. (2013). Jenaesthetics—a public dataset of paintings for aesthetic research. In *Poster workshop at the european conference on computer vision*.

Aydın, T. O., Smolic, A., & Gross, M. (2014). Automated aesthetic analysis of photographic images. *IEEE Transactions on Visualization and Computer Graphics*, *21*(1), 31–42.

Barrett, T. (2020). *Criticizing photographs: An introduction to understanding images*. Routledge.

Bhattacharya, S., Nojavanasghari, B., Chen, T., Liu, D., Chang, S. F., & Shah, M. (2013). Towards a comprehensive computational model for aesthetic assessment of videos. In *Proceedings of the 21st ACM international conference on Multimedia* (pp. 361–364).

Bianco, S., Cadene, R., Celona, L., & Napoletano, P. (2018a). Benchmark analysis of representative deep neural network architectures. *IEEE Access*, *6*, 64270–64277.

Bianco, S., Celona, L., & Schettini, R. (2018b). Aesthetics assessment of images containing faces. In *2018 25th IEEE international conference on image processing (ICIP)* (pp. 2820–2824). IEEE.

Birkhoff, G. D. (1933). *Aesthetic measure*. Cambridge, MA: Harvard University Press.

Cela-Conde, C. J., Ayala, F. J., Munar, E., Maestú, F., Nadal, M., Capó, M. A., del Río, D., López-Ibor, J. J., Ortiz, T., Mirasso, C., et al. (2009). Sex-related similarities and differences in the

neural correlates of beauty. *Proceedings of the National Academy of Sciences*, *106*(10), 3847–3852.

Chang, H., Yu, F., Wang, J., Ashley, D., & Finkelstein, A. (2016). Automatic triage for a photo series. *ACM Transactions on Graphics (TOG)*, *35*(4), 1–10.

Chang, K. Y., Lu, K. H., & Chen, C. S. (2017). Aesthetic critiques generation for photos. In *Proceedings of the IEEE international conference on computer vision* (pp. 3514–3523).

Chatterjee, A. (2003). Prospects for a cognitive neuroscience of visual aesthetics. *Bulletin of Psychology and the Arts*, *4*, 55–60. https://doi.org/10.1037/e514602010-003

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (vol. 1, pp. 539–546). https://doi.org/10.1109/CVPR.2005.202

Constantin, M. G., Kang, C., Dinu, G., Dufaux, F., Valenzise, G., & Ionescu, B. (2019). Using aesthetics and action recognition-based networks for the prediction of media memorability. In *MediaEval 2019 workshop*. France: Sophia Antipolis. https://hal.archives-ouvertes.fr/hal-02368920.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV* (vol. 1, pp. 1–2). Prague.

Cui, C., Fang, H., Deng, X., Nie, X., Dai, H., & Yin, Y. (2017). Distribution-oriented aesthetics assessment for image search. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval* (pp. 1013–1016).

Cui, C., Yang, W., Shi, C., Wang, M., Nie, X., & Yin, Y. (2020). Personalized image quality assessment with social-sensed aesthetic preference. *Information Sciences*, *512*, 780–794.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision* (pp. 288–301). Springer.

Datta, R., & Wang, J. Z. (2010). ACQUINE: aesthetic quality inference engine-real-time automatic rating of photo aesthetics. In *Proceedings of the international conference on multimedia information retrieval* (pp. 421–424).

Deng, Y., Loy, C.C., & Tang, X. (2017). Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, *34*(4), 80–106.

Deng, Y., Loy, C. C., & Tang, X. (2018). Aesthetic-driven image enhancement by adversarial learning. In *2018 ACM multimedia conference on multimedia conference* (pp. 870–878). ACM.

Dhar, S., Ordonez, V., & Berg, T. L. (2011). High level describable attributes for predicting aesthetics and interestingness. In *CVPR 2011* (pp. 1657–1664). IEEE.

Ghosal, K., Rana, A., & Smolic., A. (2019). Aesthetic image captioning from weakly-labelled photographs. In *ICCV 2019 workshop on cross-modal learning in real world*.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. Preprint. arXiv:1412.6572.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Götz-Hahn, F., Hosu, V., Lin, H., & Saupe, D. (2019). No-reference video quality assessment using multi-level spatially pooled features. Preprint. arXiv:1912.07966.

Guo, G., Wang, H., Shen, C., Yan, Y., & Liao, H. Y. M. (2018). Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression. *IEEE Transactions on Multimedia*, *20*(8), 2073–2085.

Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., & Van Gool, L. (2013). The interestingness of images. In *Proceedings of the IEEE international conference on computer vision* (pp. 1633–1640).

Hayn-Leichsenring, G. U., Lehmann, T., & Redies, C. (2017). Subjective ratings of beauty and aesthetics: correlations with statistical image properties in western oil paintings. *i-Perception*, *8*(3), 2041669517715474.

He, J., Wang, L., Zhou, W., Zhang, H., Cui, X., & Guo, Y. (2019). Viewpoint assessment and recommendation for photographing architectures. *IEEE Transactions on Visualization and Computer Graphics*, *25*(8), 2636–2649. https://doi.org/10.1109/TVCG.2018.2853751.

Hoenig, F. (2005). Defining computational aesthetics. In *Proceedings of the first eurographics conference on computational aesthetics in graphics, visualization and imaging* (pp. 13–18).

Hoffer, E., & Ailon, N. (2015). Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition* (pp. 84–92). Springer.

Hulusic, V., Valenzise, G., Provenzi, E., Debattista, K., & Dufaux, F. (2016). Perceived dynamic range of HDR images. In *IEEE int. conference on quality of multimedia experience* (pp. 1–6).

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? In *CVPR 2011* (pp. 145–152). IEEE.

ITU-R. (2012). Methodology for the subjective assessment of the quality of television pictures. ITU-R Recommendation BT.500-13.

Jaakkola, T. S., Haussler, D., et al. (1999). Exploiting generative models in discriminative classifiers. *Advances in Neural Information Processing Systems*, 487–493.

Jin, B., Segovia, M. V. O., & Süsstrunk, S. (2016a). Image aesthetic predictors based on weighted CNNs. In *IEEE international conference on image processing* (pp. 2291–2295). Phoenix, AZ, USA: IEEE.

Jin, X., Chi, J., Peng, S., Tian, Y., Ye, C., & Li, X. (2016b). Deep image aesthetics classification using inception modules and fine-tuning connected layer. In *2016 8th international conference on wireless communications & signal processing (WCSP)*, (pp. 1–6). IEEE.

Jin, X., Wu, L., Li, X., Chen, S., Peng, S., Chi, J., Ge, S., Song, C., & Zhao, G. (2018). Predicting aesthetic score distribution through cumulative Jensen-Shannon divergence. In *Thirty-second AAAI conference on artificial intelligence*.

John, L. K., Mochon, D., Emrich, O., & Schwartz, J. (2017). What's the value of a like. *Harvard Business Review*, *95*(2), 108–115.

Kairanbay, M., See, J., Wong, L. K., & Hii, Y. L. (2017). Filling the gaps: Reducing the complexity of networks for multi-attribute image aesthetic prediction. In *Proceedings of the IEEE international conference on image processing* (pp. 3051–3055). IEEE.

Kairanbay, M., See, J., & Wong, L. K. (2019). Beauty is in the eye of the beholder: Demographically oriented analysis of aesthetics in photographs. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *15*(2s), 1–21. https://doi.org/10.1145/3328993.

Kang, C., Valenzise, G., & Dufaux, F. (2019). Predicting Subjectivity in Image Aesthetics Assessment. In: 21st international workshop on multimedia signal processing (MMSP'2019), Kuala Lumpur, Malaysia. https://hal.archives-ouvertes.fr/hal-02191142.

Kang, C., Valenzise, G., & Dufaux, F. (2020). EVA: An Explainable Visual Aesthetics Dataset. In: *Joint workshop on aesthetic and technical quality assessment of multimedia and media analytics for societal trends (ATQAM/MAST'20)*. Seattle, USA: ACM Multimedia. https://hal.archives-ouvertes.fr/hal-02934292.

Kao, Y., Wang, C., & Huang, K. (2015). Visual aesthetic quality assessment with a regression model. In *IEEE international conference on image processing (ICIP)* (pp. 1583–1587). IEEE.

Kao, Y., Huang, K., & Maybank, S. (2016). Hierarchical aesthetic quality assessment using deep convolutional neural networks. *Signal Processing: Image Communication*, *47*, 500–510.

Kao, Y., He, R., & Huang, K. (2017a). Automatic image cropping with aesthetic map and gradient energy map. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1982–1986). IEEE.

Kao, Y., He, R., & Huang, K. (2017b). Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing*, *26*(3), 1482–1495.

Ke, Y., Tang, X., & Jing, F. (2006). The design of high-level features for photo quality assessment. In *IEEE int. conference on computer vision and pattern recognition (CVPR)* (vol. 1, pp. 419–426). IEEE.

Kim, W. H., Choi, J. H., & Lee, J. S. (2020). Objectivity and subjectivity in aesthetic quality assessment of digital photographs. *IEEE Transactions on Affective Computing*, *11*(3), 493–506. https://doi.org/10.1109/TAFFC.2018.2809752.

Kirk, U., Skov, M., Christensen, M. S., & Nygaard, N. (2009). Brain correlates of aesthetic expertise: a parametric fMRI study. *Brain and Cognition*, *69*(2), 306–315.

Kong, S., Shen, X., Lin, Z., Mech, R., & Fowlkes, C. (2016). Photo aesthetics ranking network with attributes and content adaptation. In *European conference on computer vision* (pp. 662–679). Springer.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Kucer, M., Loui, A. C., & Messinger, D. W. (2018). Leveraging expert feature knowledge for predicting image aesthetics. *IEEE Transactions on Image Processing*, *27*(10), 5100–5112.

Kuzovkin, D. (2019). Assessment of photos in albums based on aesthetics and context. Theses, Université Rennes 1. https://hal.inria.fr/tel-02345620.

Kuzovkin, D., Pouli, T., Cozot, R., Meur, O. L., Kervec, J., & Bouatouch, K. (2017). Context-aware clustering and assessment of photo collections. In *Proceedings of the symposium on computational aesthetics* (pp. 1–10)

Lakhal, S., Darmon, A., Bouchaud, J. P., & Benzaquen, M. (2020). Beauty and structural complexity. *Physical Review Research*, *2*(2), 022058.

Lebreton, P., Raake, A., & Barkowsky, M. (2016). Evaluation of aesthetic appeal with regard of user's knowledge. *Electronic Imaging*, *2016*(16), 1–6.

Li, C., Loui, A. C., & Chen, T. (2010). Towards aesthetics: A photo quality assessment and photo selection system. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 827–830).

Li, J., Barkowsky, M., & Le Callet, P. (2013). Boosting paired comparison methodology in measuring visual discomfort of 3dtv: performances of three different designs. In *Stereoscopic displays and applications XXIV* (vol. 8648, p. 86481V). International Society for Optics and Photonics.

Ling, S., Wang, J., Huang, W., Guo, Y., Zhang, L., Jing, Y., & Le Callet, P. (2020). A subjective study of multi-dimensional aesthetic assessment for mobile game image. In *Proceedings of the 1st workshop on quality of experience (QoE) in visual multimedia applications* (pp. 47–53).

Liu, W., & Wang, Z. (2017). A database for perceptual evaluation of image aesthetics. In *2017 IEEE international conference on image processing (ICIP)* (pp. 1317–1321). https://doi.org/10.1109/ICIP.2017.8296495.

Liu, D., Puri, R., Kamath, N., & Bhattacharya, S. (2020). Composition-aware image aesthetics assessment. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*.

Lo, K. Y., Liu, K. H., & Chen, C. S. (2012). Intelligent photographing interface with on-device aesthetic quality assessment. In *Asian conference on computer vision* (pp. 533–544). Springer.

Lowe, D. G.: Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision* (vol. 2, pp. 1150–1157). IEEE (1999)

Lu, X., Lin, Z., Jin, H., Yang, J., & Wang, J. Z. (2014). RAPID: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 457–466). ACM.

Lu, P., Peng, X., Li, R., & Wang, X. (2015a). Towards aesthetics of image: a bayesian framework for color harmony modeling. *Signal Processing: Image Communication*, *39*, 487–498.

Lu, X., Lin, Z., Jin, H., Yang, J., & Wang, J. Z. (2015b). Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, **17**(11), 2021–2034. https://doi.org/10.1109/TMM.2015.2477040.

Lu, X., Lin, Z., Shen, X., Mech, R., & Wang, J. Z. (2015c). Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *Proceedings of the IEEE international conference on computer vision* (pp. 990–998)

Lu, P., Peng, X., Zhu, X., & Li, R. (2016). An EL-LDA based general color harmony model for photo aesthetics assessment. *Signal Processing*, *120*, 731–745.

Luo, Y., & Tang, X. (2008). Photo and video quality evaluation: Focusing on the subject. In *European conference on computer vision* (pp. 386–399). Springer.

Luo, W., Wang, X., & Tang, X. (2011). Content-based photo quality assessment. In *2011 international conference on computer vision* (pp. 2206–2213). IEEE.

Ma, S., Liu, J., & Wen Chen, C. (2017). A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4535–4544).

Machado, P., & Cardoso, A. (1998). Computing aesthetics. In *Brazilian symposium on artificial intelligence* (pp. 219–228). Springer.

Mai, L., Le, H., Niu, Y., & Liu, F. (2011). Rule of thirds detection from photograph. In *2011 IEEE international symposium on Multimedia* (pp. 91–96).

Mai, L., Le, H., Niu, Y., Lai, Y. C., & Liu, F. (2012). Detecting rule of simplicity from photos. In *Proceedings of the 20th ACM international conference on Multimedia* (pp. 1149–1152)

Mai, L., Jin, H., & Liu, F. (2016). Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 497–506).

Maître, H. (2018). Qu'est-ce qu'une belle photo? Essai sur l'esthétique en photographie numérique. https://hal.archives-ouvertes.fr/hal-01864135/.

Marchesotti, L., Perronnin, F., & Meylan, F. (2013). Learning beautiful (and ugly) attributes. In *BMVC* (vol. 7, pp. 1–11).

Marchesotti, L., Perronnin, F., Larlus, D., & Csurka, G. (2011). Assessing the aesthetic quality of photographs using generic image descriptors. In *IEEE international conference on computer vision* (pp. 1784–1791). IEEE.

Moon, P., & Spencer, D. (1944). Geometric formulation of classical color harmony. *Journal of the Optical Society of America (1917-1983)*, *34*(1), 46.

Murray, N., & Gordo, A. (2017). A deep architecture for unified aesthetic prediction. Preprint. arXiv:1708.04890.

Murray, N., Marchesotti, L., & Perronnin, F. (2012). AVA: a large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, (pp. 2408–2415). IEEE.

Park, T. S., & Zhang, B. T. (2015). Consensus analysis and modeling of visual aesthetic perception. *IEEE Transactions on Affective Computing*, *6*(3), 272–285.

Park, K., Hong, S., Baek, M., & Han, B. (2017). Personalized image aesthetic quality assessment by joint regression and ranking. In *2017 IEEE winter conference on applications of computer vision (WACV)* (pp. 1206–1214). IEEE.

Pele, O., & Werman, M. (2010). The quadratic-chi histogram distance family. In *European conference on computer vision* (pp. 749–762). Springer.

Perez-Ortiz, M., Mikhailiuk, A., Zerman, E., Hulusic, V., Valenzise, G., & Mantiuk, R. (2019). From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing*, *29*, 1139–1151. https://doi.org/10.1109/TIP.2019.2936103. https://hal.archives-ouvertes.fr/hal-02400863.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, *8*(4), 364–382.

Redi, M., Rasiwasia, N., Aggarwal, G., & Jaimes, A. (2015). The beauty of capturing faces: Rating the quality of digital portraits. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition* (vol. 1, pp. 1–8). IEEE.

Redi, M., Crockett, D., Manovich, L., & Osindero, S. (2016). What makes photo cultures different? In *Proceedings of the 24th ACM international conference on Multimedia* (pp. 287–291).

Ren, J., Shen, X., Lin, Z., Mech, R., & Foran, D. J. (2017). Personalized image aesthetics. In *Proceedings of the IEEE international conference on computer vision* (pp. 638–647).

Ribeiro, F., Florêncio, D., Zhang, C., & Seltzer, M. (2011). Crowdmos: An approach for crowdsourcing mean opinion score studies. In *IEEE international conference on acoustics, speech and signal processing* (pp. 2416–2419). IEEE.

Rigau, J., Feixas, M., & Sbert, M. (2008). Informational aesthetics measures. *IEEE Computer Graphics and Applications*, *28*(2), 24–34.

Rosenblum, N. (2008). *A world history of photography*. New York, USA: Abbeville Press.

San Pedro, J., Yeh, T., & Oliver, N. (2012). Leveraging user comments for aesthetic aware image search reranking. In *Proceedings of the 21st international conference on World Wide Web* (pp. 439–448).

Schifanella, R., Redi, M., & Aiello, L. M. (2015). An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. In *ICWSM'15: Proceedings of the 9th AAAI international conference on weblogs and social media*. AAAI.

Schwarz, K., Wieschollek, P., & Lensch, H. P. A. (2018). Will people like your image? learning the aesthetic space. In *2018 IEEE winter conference on applications of computer vision (WACV)* (pp. 2048–2057). https://doi.org/10.1109/WACV.2018.00226.

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Sheng, K., Dong, W., Huang, H., Ma, C., & Hu, B. G. (2018a). Gourmet photography dataset for aesthetic assessment of food images. In *SIGGRAPH Asia 2018 technical briefs* (pp. 1–4).

Sheng, K., Dong, W., Ma, C., Mei, X., Huang, F., & Hu, B. G. (2018b). Attention-based multi-patch aggregation for image aesthetic assessment. In *Proceedings of the 26th ACM international conference on Multimedia* (pp. 879–886).

Siahaan, E., Hanjalic, A., & Redi, J. (2016). A reliable methodology to collect ground truth data of image aesthetic appeal. *IEEE Transactions on Multimedia*, *18*(7), 1338–1350. https://doi.org/10.1109/TMM.2016.2559942.

Simond, F., Arvanitopoulos, N., & Süsstrunk, S. (2015). Image aesthetics depends on context. In *IEEE international conference on image processing (ICIP)* (pp. 3788–3792). IEEE.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. Preprint. arXiv:1409.1556.

Song, Y., Redi, M., Vallmitjana, J., & Jaimes, A. (2016). To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 659–668).

Spape, M., Davis, K., Kangassalo, L., Ravaja, N., Sovijarvi-Spape, Z., & Ruotsalo, T. (2021). Brain-computer interface for generating personally attractive images. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2021.3059043.

Stollenga, M. F., Masci, J., Gomez, F., & Schmidhuber, J. (2014). Deep networks with internal selective attention through feedback connections. In *Proceedings of the 27th international conference on neural information processing systems* (vol. 2, pp. 3545–3553).

Suchecki, M., & Trzciski, T. (2017). Understanding aesthetics in photography using deep convolutional neural networks. In *2017 Signal processing: Algorithms, architectures, arrangements, and applications (SPA)* (pp. 149–153). https://doi.org/10.23919/SPA.2017.8166855.

Sun, W., Chao, T., Kuo, Y., & Hsu, W. H. (2017). Photo filter recommendation by category-aware aesthetic learning. *IEEE Transactions on Multimedia*, *19*(8), 1870–1880. https://doi.org/10.1109/TMM.2017.2688929.

Talebi, H., & Milanfar, P. (2018). NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, *27*(8), 3998–4011.

Tang, X., Luo, W., & Wang, X. (2013). Content-based photo quality assessment. *IEEE Transactions on Multimedia*, *15*(8), 1930–1943.

Tifentale, A., & Manovich, L. (2018). Competitive photography and the presentation of the self. In *Exploring the selfie* (pp. 167–187). Springer.

Wang, W., Zhao, M., Wang, L., Huang, J., Cai, C., & Xu, X. (2016). A multi-scene deep learning model for image aesthetic evaluation. *Signal Processing: Image Communication*, *47*, 511–518.

Wang, Z., Liu, D., Chang, S., Dolcos, F., Beck, D., & Huang, T. (2017). Image aesthetics assessment using Deep Chatterjee's machine. In *International joint conference on neural networks (IJCNN)* (pp. 941–948). https://doi.org/10.1109/IJCNN.2017.7965953.

Wang, W., Yang, S., Zhang, W., & Zhang, J. (2019). Neural aesthetic image reviewer. *IET Computer Vision*, *13*(8), 749–758.

Wu, O., Hu, W., & Gao, J. (2011). Learning to predict the perceived visual quality of photos. In *International conference on computer vision* (pp. 225–232). https://doi.org/10.1109/ICCV.2011.6126246.

Xu, M., Chen, F., Li, L., Shen, C., Lv, P., Zhou, B., & Ji, R. (2018). Bio-inspired deep attribute learning towards facial aesthetic prediction. *IEEE Transactions on Affective Computing*, 227–238.

Xu, Y., Zhang, N., Wei, P., Sang, G., Li, L., & Yuan, F. (2020). Deep neural framework with visual attention and global context for predicting image aesthetics. *IEEE Access*, 1–1. https://doi.org/10.1109/ACCESS.2020.3015060.

Ye, P., & Doermann, D. (2014). Active sampling for subjective image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4249–4256).

Yeh, H. H., Yang, C. Y., Lee, M. S., & Chen, C. S. (2013). Video aesthetic quality assessment by temporal integration of photo-and motion-based features. *IEEE Transactions on Multimedia*, *15*(8), 1944–1957.

Yu, J., Cui, C., Geng, L., Ma, Y., & Yin, Y. (2019). Towards unified aesthetics and emotion prediction in images. In *2019 IEEE international conference on image processing (ICIP)* (pp. 2526–2530). https://doi.org/10.1109/ICIP.2019.8803388.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer.

Zerman, E., Hulusic, V., Valenzise, G., Mantiuk, R., & Dufaux, F. (2018). The relation between MOS and pairwise comparisons and the importance of cross-content comparisons. In *Human vision and electronic imaging conference, IS&T international symposium on electronic imaging*, Burlingame, USA. https://hal.archives-ouvertes.fr/hal-01654133.

Zerman, E., Rana, A., & Smolic, A. (2019). Colornet-estimating colorfulness in natural images. In *IEEE international conference on image processing* (pp. 3791–3795). IEEE.

Zhang, F., Wang, M., & Hu, S. (2013). Aesthetic image enhancement by dependence-aware object recomposition. *IEEE Transactions on Multimedia*, *15*(7), 1480–1490. https://doi.org/10.1109/TMM.2013.2268051.

Zhang, L., Gao, Y., Zhang, C., Zhang, H., Tian, Q., & Zimmermann, R. (2014). Perception-guided multimodal feature fusion for photo aesthetics assessment. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 237–246).

Zhang, C., Zhu, C., Xu, X., Liu, Y., Xiao, J., & Tillo, T. (2018). Visual aesthetic understanding: Sample-specific aesthetic classification and deep activation map visualization. *Signal Processing: Image Communication*, *67*, 12–21.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921–2929).

Zuckert, R. (2007). *Kant on Beauty and Biology: An Interpretation of the 'Critique of Judgment'*. Cambridge University Press.