



HAL
open science

Image Aesthetic Quality Assessment Based on Deep Neural Networks

Chen Kang

► **To cite this version:**

Chen Kang. Image Aesthetic Quality Assessment Based on Deep Neural Networks. Image Processing [eess.IV]. Université Paris-Saclay, 2020. English. NNT : 2020UPASG004 . tel-03159861

HAL Id: tel-03159861

<https://theses.hal.science/tel-03159861v1>

Submitted on 4 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Image Aesthetic Quality Assessment Based on Deep Neural Networks

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580 sciences et technologies de
l'information et de la communication (STIC)
Spécialité de doctorat: Traitement du signal et des images
Unité de recherche: Université Paris-Saclay, CNRS, CentraleSupélec,
Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France.
Réfèrent: Faculté des sciences d'Orsay

**Thèse présentée et soutenue en visioconférence totale,
le 14 Décembre 2020, par**

Chen KANG

Composition du jury:

Sylvie LE HEGARAT Professeure des Universités, Université Paris-Saclay	Présidente
Rémi COZOT Professeur des Universités, Université du Littoral Côte d'Opale	Rapporteur & Examinateur
Lu ZHANG Maîtresse de Conférence (HDR), INSA Rennes	Rapportrice & Examinatrice
Aladine CHETOUANI Maître de Conférence, Université d'Orléans	Examinateur
Patrick LE CALLET Professeur des Universités, Université de Nantes	Examinateur
Giuseppe VALENZISE CR CNRS, L2S, CNRS, CentraleSupélec	Directeur de thèse
Frédéric DUFAUX DR CNRS, L2S, CNRS, CentraleSupélec	Co-directeur de thèse

Abstract

With the development of capture devices and the Internet, people access to an increasing amount of images. Assessing visual aesthetics has important applications in several domains, from image retrieval and recommendation to enhancement.

Image aesthetic quality assessment aims at determining how beautiful an image looks to human observers. Many problems in this field are not studied well, including the subjectivity of aesthetic quality assessment, explanation of aesthetics and the human-annotated data collection. Conventional image aesthetic quality prediction aims at predicting the average score or aesthetic class of a picture. However, the aesthetic prediction is intrinsically subjective, and images with similar mean aesthetic scores/class might display very different levels of consensus by human raters. Recent work has dealt with aesthetic subjectivity by predicting the distribution of human scores, but predicting the distribution is not directly interpretable in terms of subjectivity, and might be sub-optimal compared to directly estimating subjectivity descriptors computed from ground-truth scores. Furthermore, labels in existing datasets are often noisy, incomplete or they do not allow more sophisticated tasks such as understanding why an image looks beautiful or not to a human observer.

In this thesis, we first propose several measures of subjectivity, ranging from simple statistical measures such as the standard deviation of the scores, to newly proposed descriptors inspired by information theory. We evaluate the prediction performance of these measures when they are computed from predicted score distributions and when they are directly learned from ground-truth data. We find that the latter strategy provides in general better results. We also use the subjectivity to improve predicting aesthetic scores, showing that information theory inspired subjectivity measures perform better than statistical method inspired subjectivity measures.

Then, we propose an Explainable Visual Aesthetics (EVA) dataset, which contains 4070 images with at least 30 votes per image. EVA has been crowd-sourced using a more disciplined approach inspired by quality assessment best practices. It also offers additional features, such as the degree of difficulty in assessing the aesthetic score, rating for 4 complementary aesthetic attributes, as well as the relative importance of each attribute to form aesthetic opinions. The publicly available dataset is expected to contribute to future research on understanding and predicting visual quality aesthetics.

Additionally, we studied the explainability of image aesthetic quality assessment. A statistical analysis on EVA demonstrates that the collected attributes and relative importance can be linearly combined to explain effectively the overall aesthetic mean opinion scores. We found

subjectivity has a limited correlation to average personal difficulty in aesthetic assessment, and the subject's region, photographic level and age affect the user's aesthetic assessment significantly.

Key word: Aesthetics, Image/Video Processing, Machine Learning, Neural Network, Visual Quality Assessment

Synthèse en français

Avec le développement des dispositifs de capture et d'Internet, les utilisateurs ont accès à une quantité croissante d'images. L'évaluation de l'esthétique visuelle a des applications importantes dans plusieurs domaines, de la recherche et de la recommandation d'images à leur amélioration. Les prédictors récents de la qualité esthétique sont basés sur les données et exploitent la disponibilité de grands ensembles de données annotées pour entraîner des modèles de prédiction précis.

L'évaluation de la qualité esthétique des images vise à déterminer la beauté d'une image pour les observateurs humains. Chercheurs ont travaillé sur la prédiction de l'esthétique générale des photos et construit différentes bases de données, mais certains problèmes dans ce domaine ne sont pas bien étudiés, notamment la subjectivité de l'évaluation de la qualité esthétique, l'explication de l'esthétique et la collecte de données annotées par des humains. Les méthodes conventionnelles pour la prédiction de la qualité esthétique des images visent à prédire le score moyen ou la classe esthétique d'une image. Cependant, la prédiction esthétique est intrinsèquement subjective, et des images ayant des notes moyennes ou des classes esthétiques similaires peuvent présenter des niveaux de consensus très différents selon les évaluateurs humains. Des travaux récents ont traité de la subjectivité esthétique en prédisant la distribution des scores humains, mais la prédiction de la distribution n'est pas directement interprétable en termes de subjectivité, et pourrait être sous-optimale par rapport à l'estimation directe des descripteurs de subjectivité calculés à partir de scores de vérité terrain. En outre, les étiquettes des bases de données existantes sont souvent bruitées, incomplètes ou ne permettent pas d'effectuer des tâches plus sophistiquées, comme comprendre pourquoi une image est belle ou non pour un observateur humain.

Dans cette thèse, nous proposons d'abord plusieurs mesures de la subjectivité, allant de simples mesures statistiques telles que l'écart-type des scores, à des nouveaux descripteurs inspirés de la théorie de l'information. Nous évaluons la performance de prédiction de ces mesures lorsqu'elles sont calculées à partir de distributions de scores prédites et lorsqu'elles sont directement apprises à partir de données de vérité terrain. Nous constatons que cette dernière stratégie donne en général de meilleurs résultats. Nous utilisons également la subjectivité pour améliorer la prédiction des scores esthétiques, en montrant que les mesures de subjectivité inspirées par la théorie de l'information donnent de meilleurs résultats que les mesures de subjectivité inspirées par la statistique.

Ensuite, nous proposons une base de données appelée «Esthétique Visuelle Explicable» (EVA), qui contient 4070 images avec au moins 30 votes par image. Les votes dans EVA

ont été obtenus en ligne en utilisant une approche plus disciplinée inspirée des meilleures pratiques d'évaluation de la qualité. La base offre également des caractéristiques supplémentaires, telles que le degré de difficulté de l'évaluation de la note esthétique, la notation de 4 attributs esthétiques complémentaires, ainsi que l'importance relative de chaque attribut dans la formation d'une opinion esthétique. Nous fournissons également les détails de la conception de l'enquête, le choix des utilisateurs et des images, la méthode de collecte des données et donnons un résumé des informations collectées. La base de données surmonte les étiquettes bruitées causées par une description de tâche ambiguë, des sujets limités et un manque de difficulté et d'attributs bien étiquetés par des humains. La base de données a été mise à disposition du public et devrait contribuer aux futures recherches sur la compréhension et la prévision de la qualité esthétique visuelle.

En outre, nous avons étudié l'explicabilité de l'évaluation de la qualité esthétique des images. Une analyse statistique sur EVA démontre que les attributs collectés et leur importance relative peuvent être combinés de manière linéaire pour expliquer efficacement les scores moyens globaux des opinions esthétiques. Nous avons constaté que la subjectivité a une corrélation limitée avec la difficulté personnelle moyenne de l'évaluation esthétique, et que la provenance géographique, le niveau photographique et l'âge du sujet affectent de manière significative l'évaluation esthétique de l'utilisateur.

Mots clés: Esthétique, Traitement des Images / Vidéos, Apprentissage Automatique, Réseau Neuronal, Évaluation de la Qualité Visuelle

Acknowledgements

Time flies. By spending three years in the Paris-Saclay area, I explored many towns and forests; I experienced the European culture and French society; I learned how to think and how to work. Moreover, I met many talented and kind people.

First, I would like to thank Dr. Lu ZHANG and Prof. Rémi COZOT for reviewing my thesis and giving me suggestions to ensure its quality. I would also like to thank other admirable jury members, Dr. Aladine CHETOUANI, Prof. Patrick LE CALLET and Prof. Sylvie LE HEGARAT, for examining my thesis and participating in my defence in the COVID-19 difficulty.

I would express my sincere thanks to my supervisors Dr. Giuseppe VALENZISE and Dr. Frédéric DUFAUX for their patient guidance and professional suggestions. It is my honour to be one of their students. Dr. VALENZISE not only teaches me and discusses with me in research, but also takes care of me and teaches me in life. His passion to desserts opened a new gate for me. Dr. DUFAUX is always available for my questions, and his words is like a light in the mist.

My deep appreciation goes for the volunteers of the EVA datasets, without whom the work would not be able to completed. I would also like to thank the good working partner Mihai Gabriel CONSTANT in memorability prediction.

My thanks with sincerity also goes to colleagues and friends in Laboratoire des Signaux et Système, CentraleSupélec and Télécom Paris: Prof. Michel KIEFFER, Dr. Pierre DUHAMEL, Prof. Gilles DUC, Ms. Maryvonne GIRON, Ms. Stéphanie DOUESNARD, Ms. Catherine NIZERY, Mr. José FONSECA, Mr. Thomas CUIDU, Mr. Christian DAVID, Ms. Maryelle FRADIN, for offering help and being kind to me; Dr. Li WANG, Maurice QUACH, Dr. Florent CHIARONI, Dr. Emin ZERMAN, Dr. Maxim KARPUSHIN, Dr. Aakanksha RANA, Dr. Shuo ZHENG, Milan STEPANOV, Abhishek GOSWAMI, for inspiring and sharing in research; Siyang LIU, Junjie YANG, Xiaoxia ZHANG, Xuewen QIAN, Dr. Xiaojun XI, Dr. Jian SONG, Dr. Weichao LIANG, Dr. Kai WAN, Li HUANG, Dr. Siqi WANG, Dr. Chao ZHANG, Hang ZOU, Yifei SUN, Dr. Chi JIN, Dr. Bowen YI, Dr. Zhenyu LIAO, Junbo TAN, Baojie LI, Peipei RAN, Dr. Shanshan WANG, Jiang LIU, Dr. Yanqiao HOU, Dung, Thanh, Saman, Fadil, Adel, Incha, Kuba, Sangwoo, et al., for the great time together.

I am also grateful for other friends for spending happy time together. Zi YE, Yingchao DOU, Shuanglin GUO, Dr. Xiaoxia FENG, Chuang YU, Qiqi HOU, Yaxiong WANG, Dr. Sibao CHENG, Dr. Rui LI, Dr. Renee BELL, Basile DUFOUR, Ning GUO, Chengzhang ZHONG, Jinjing ZHU, Lili LIU, and many others shared supports, inspirations, and discussions with me. Gaoya YANG and Qiudan CHEN discussed about visual aesthetics with me and provided professional suggestions. I am lucky to have Dr. Zihan GENG, Dr. Xiaoyan LIU, Jiajie CHENG, Xinrui JI,

Binghao JIA, and Ning ZHANG there for many years to share happy and frustrations. Uncle Dr. CHEN, Aunt GUO, Sophie, Xi and Kevin always care about me and provide a cozy place in festivals.

Most importantly, I want to devote this thesis to my husband Dr. Tianyu ZHENG and our parents. Our parents never lose confidence in our potentials and always encourage us to broaden the horizon. Thanks to Tianyu for being caring and loving many years, and coming to Île-de-France together to pursue our PhDs .

At last, my deep thanks shall give to China Scholarship Council (CSC) for funding me to work on this challenging topic and to explore the world.

Contents

Abstract	i
Synthèse en français	iii
Acknowledgements	v
List of Figures	ix
List of Tables	xii
1 Introduction and Overview	2
1.1 Image Aesthetic Quality Assessment	2
1.2 Tackled Problems of the Thesis	5
1.2.1 Subjectivity in Aesthetic Image Assessment	5
1.2.2 Explanation of Computational Aesthetics	6
1.2.3 Human-annotated Data Collection	8
1.3 Objectives of the Thesis	9
1.4 Structure of the Thesis	10
2 Related Work	11
2.1 Image Aesthetic Quality Prediction	11
2.1.1 Aesthetic Assessment Pipeline and Deep Neural Networks	11
2.1.2 Background of Predicting Aesthetic quality from Images	14
2.2 Predicting Aesthetic Subjectivity	20
2.3 Explanation Attempts in Images Aesthetics	23
2.4 Existing Dataset Overview	25
3 Modeling and Predicting Subjectivity in Aesthetics	33
3.1 Introduction	33
3.2 Proposed Measures of Subjectivity	34

3.3	Subjectivity Prediction Framework	36
3.4	Experimental Results	37
3.4.1	Experimental Setup	37
3.4.2	Performance Indicators	39
3.4.3	Experiment Results	40
3.5	Improvement of Mean Aesthetic Score Prediction	42
3.5.1	Method	42
3.5.2	Experiment and Result	43
3.6	Conclusion	45
4	EVA: an Explainable Visual Aesthetics dataset	49
4.1	Work Flow, Platform and Experiment Settings	49
4.2	Survey Design	57
4.3	Image Selection	60
4.4	Data Quality Control and Cleaning Procedure	62
4.5	Data Summary	66
4.6	Conclusion	70
5	Aesthetics and Attributes	75
5.1	Analysis of Aesthetic Score and Attributes	75
5.1.1	Relation between Attributes and Aesthetic Score	77
5.1.2	Difficulty and Subjectivity in the Aesthetic Evaluation	79
5.1.3	Relation between Attributes and Subjectivity	80
5.2	Aesthetics and User Characteristics	83
5.3	Conclusion	88
6	Conclusion and Perspective	93
6.1	Summary of the Contributions	93
6.2	Perspectives	94
A	Aesthetics in Media Memorability Prediction	97
A.1	Introduction	97
A.2	Approach	98
A.2.1	Aesthetics networks	98
A.2.2	Action recognition networks	98
A.2.3	Late fusion	99
A.3	Experimental results	100

A.3.1 Results on the devset	100
A.3.2 Results on the testset	101
A.4 Conclusions	101
B Author's publications during 3 years of PhD	102
Bibliography	103

List of Figures

1.1	Examples of aesthetically good and bad images	3
1.2	Example of aesthetic subjectivity for two images of the AVA dataset. The two images, displayed in the top-left panels, have similar mean score but different distribution of aesthetic judgments given by human raters, shown in the histograms on the top-right panels. The tables report the mean score and standard deviation of the votes.	7
2.1	Network architecture of AlexNet [56]	13
2.2	Network architecture of VGG-16 [96]	13
2.3	Network architecture of ResNet-34 [37]	14
2.4	Scope of image aesthetic quality prediction	14
2.5	Scope of improving aesthetic value prediction methods	16
2.6	Distribution of predicted standard deviation and ground-truth's standard deviation [101].	21
2.7	Images with similar mean scores share different histograms, variance, skewness and kurtosis [44].	22
2.8	An aesthetic signature example with the 5 proposed visual attributes [2].	25
2.9	Example images in different style categories of AVA dataset [77]. They showed their classification of images in good (green) and bad (red)	27
2.10	Example images of dataset in [10].	28
2.11	Examples of AADB dataset [55].	29
2.12	Example images from the Waterloo IAA dataset [65]. From the top to the bottom row represent Animal, Architecture/City Scenes, Human, Natural Scene, Still Object images respectively.	30
3.1	The two score distributions have the same entropy, but the one on the right has a higher degree of subjectivity.	35

3.2	Measures that compactly describe subjectivity based on the two score distributions in Figure 1.2 respectively.	36
3.3	Subjectivity Prediction Framework. In the indirect prediction framework, an aesthetic score distribution is estimated first, and subjectivity measures are computed over it. We compare this approach with directly predicting subjectivity computed on ground-truth distributions (b).	38
3.4	Average predicted score distribution vs. ground-truth score distribution over the test set, for the three considered state-of-the-art distribution prediction methods. Notice that for all of them, the average predicted distribution is shifted compared to the original.	46
3.5	Network structures of training and testing	47
3.6	AVA subjectivity distribution. Higher SD and MAD indicates a lower consensus, and higher DUD and MED means a higher consensus.	48
3.7	Relationship between the square of standard deviation and MOS in AVA dataset.	48
4.1	General Flow of Survey	50
4.2	Training's screenshot, part 1.	51
4.3	Training's screenshot, part 2.	52
4.4	Training's screenshot, part 3.	53
4.5	Training's screenshot, part 4.	54
4.6	Training's screenshot, part 5.	55
4.7	Synchronization of two web hosts	57
4.8	Voting page's screenshot on mobile devices. When the user pulls the bar in the first question, there will be a score under the bar. If one of the 4 questions is not answered, the user cannot trigger the "Save" button, and there will be an alert.	59
4.9	Images in AVA dataset with MOS under 4	61
4.10	AVA score distribution for the images in EVA.	62
4.11	Image examples in each category for EVA	63
4.12	Each user's general aesthetic assessment's mean score, standard deviation and number of votes. The red frames show the suspected outliers.	66
4.13	Distribution of vote number of users whose voting general aesthetic scores' standard deviation is smaller than 0.1	67
4.14	Number of votes for each image	67
4.15	Statistics of votes in EVA dataset: (a) Gender (b) Region (c) Visual Status (d) Photography Experience (e) Age (f) Device	68
4.16	Distribution of Mean Opinion Score (MOS)	69

4.17 Distribution of Standard Deviation (STD) of scores	69
4.18 Distribution of user's vote number	70
4.19 Distribution of attributes. (a) Light and color (b) Composition and depth (c) Quality (d) Semantic	71
4.20 Average probability for one attribute to affect the overall aesthetic judgment.	72
4.21 Average distribution of attributes importance per content category: (1) animals (2) architectures and city scenes (3) human (4) natural and rural scenes (5) still life (6) other	73
4.22 Comparison between AVA's MOS and EVA's MOS	74
4.23 QQ-plot of AVA's MOS and EVA's MOS	74
4.24 Comparison between AVA's MOS and EVA's MOS	74
5.1 Relationship between image's MOS and each attribute's mean opinion score.	77
5.2 Average distribution of attributes importance per content category: (1) animals (2) architectures and city scenes (3) human (4) natural and rural scenes (5) still life (6) other	81
5.3 QQ plot of different genders	84
5.4 Distribution of different region's user's mean score	85
5.5 QQ plot of different regions	86
5.6 Distribution of different photographic level's user's mean score	87
5.7 QQ plot of beginners and intermediate subjects	87
5.8 Distribution of different eye status' user's mean score	88
5.9 QQ plot of different eye status' user's mean score. Non-colourblind includes nor- mal subjects and the ones with glasses; colourblind includes only having colour- blindness and having colourblindness and glasses at the same time.	89
5.10 QQ plot of different eye status' user's mean colour and light score	90
5.11 Distribution of different ages' user's mean votes	90
5.12 QQ plots of different ages' user's mean votes	91
A.1 The diagram of the proposed solution.	99

List of Tables

2.1	Information utilised in general image aesthetics prediction	16
2.2	Datasets Inferring Aesthetic Information Indirectly	26
2.3	Overview of datasets that infers aesthetic quality directly without album context. ✓ means "yes", × means "not designed" or "not provided". "D" means the aes- thetic distributions are provided or can be computed from the public dataset files, "S" is for the aesthetic mean score, "B" is for binary labels. "Website" is for "photo-sharing website". "Lab" is for "lab environment".	32
3.1	Pearson's Linear Correlation Coefficient (PLCC)	40
3.2	Spearman's Rank-Order Correlation Coefficient (SROCC)	40
3.3	Mean Absolute Error (MAE)	40
3.4	Mean Relative Absolute Error (MRAE)	41
3.5	Predicting Aesthetic Score with Weights	43
4.1	Level of photographic experience	49
4.2	Comparison between EVA and milestone datasets ✓ means "yes", × means "not designed" or "not provided".	72
5.1	Correlation between Mean Score of Attributes and Mean Opinion Score (MOS). STD denotes the standard deviation of the global aesthetic scores per image. The numbers are SROCC/PLCC.	76
5.2	Correlation between Subjectivity Measurements, Mean Opinion Score (MOS), Difficulty and Mean Score of Attributes. The numbers are SROCC/PLCC.	82
A.1	Results of the proposed runs (preliminary experiments on <i>devset</i> , and official results on <i>testset</i>).	99

Chapter 1

Introduction and Overview

With the development of acquisition devices and online communication, an enormous quantity of images are created, uploaded and shown to people in daily life and work. Because of the quantity of visual data and time-consuming demand, the study of computer vision and visual processing algorithms is increasing. Besides objective tasks like recognition of objects, cars and people's faces, detection of pedestrians and objects, reconstruction of architectures from 2D images, point cloud compression, etc., many subjective tasks attract attentions as well. Among them, the subjective task of selecting a preferred group of photos from a large dataset is used in many areas such as retrieval, recommendation, enhancement, summarization, memorability, etc. Taking a digital photo has become easy for all ages, but a good quality photo does not mean a beautiful photo.

Sometimes users use photo editing softwares like Photoshop, Lightroom, and CorelDRAW Photo-Paint, or mobile applications like VSCO, Meitu, Snapseed, etc., to meet their aesthetic expectations. When people retrieve an image or a sketch, there are usually a list of correct results that are similar to the query in shape and pattern. Ranking their aesthetically better images higher leads to a positive user experience.

In general, choosing and editing a beautiful image that the majority agrees on its aesthetic value is time-consuming and difficult for non-professional people. Using algorithms to do it can save resources.

1.1 Image Aesthetic Quality Assessment

Aesthetic means relating to the enjoyment or study of beauty in the *Cambridge Dictionary*, and aesthetics is a branch of philosophy that deals with the nature of beauty and taste [1]. Beauty means the quality of being pleasing, especially to look at, or someone or something

that gives great pleasure. In one aspect, this definition shows that beauty is a kind of quality of someone or something. In the other aspect, as a subjective task, "beauty is in the eye of the beholder" - there is an intrinsic difficulty of finding a "standard answer". The definition and the goal of the subjective tasks are usually controversial across different fields. In philosophy, from Plato to Kant, from objectivist to subjectivist, what is beauty and what is beautiful does not have a conclusion. As philosophers used different European languages in different epochs, the interpretation to "beauty" got different emphasised aspects and changed progressively. Also, in contemporary time, when the aesthetics was introduced to other cultures, local philosophies gave other interpretations of beauty into aesthetics.

In the field of multimedia, the goal of image aesthetic quality assessment is to determine how beautiful an image looks to human observers. For example, images shown in Figure 1.1(a) and 1.1(b) are considered beautiful by most people, and Figure 1.1(c) and 1.1(d) are considered as not beautiful.

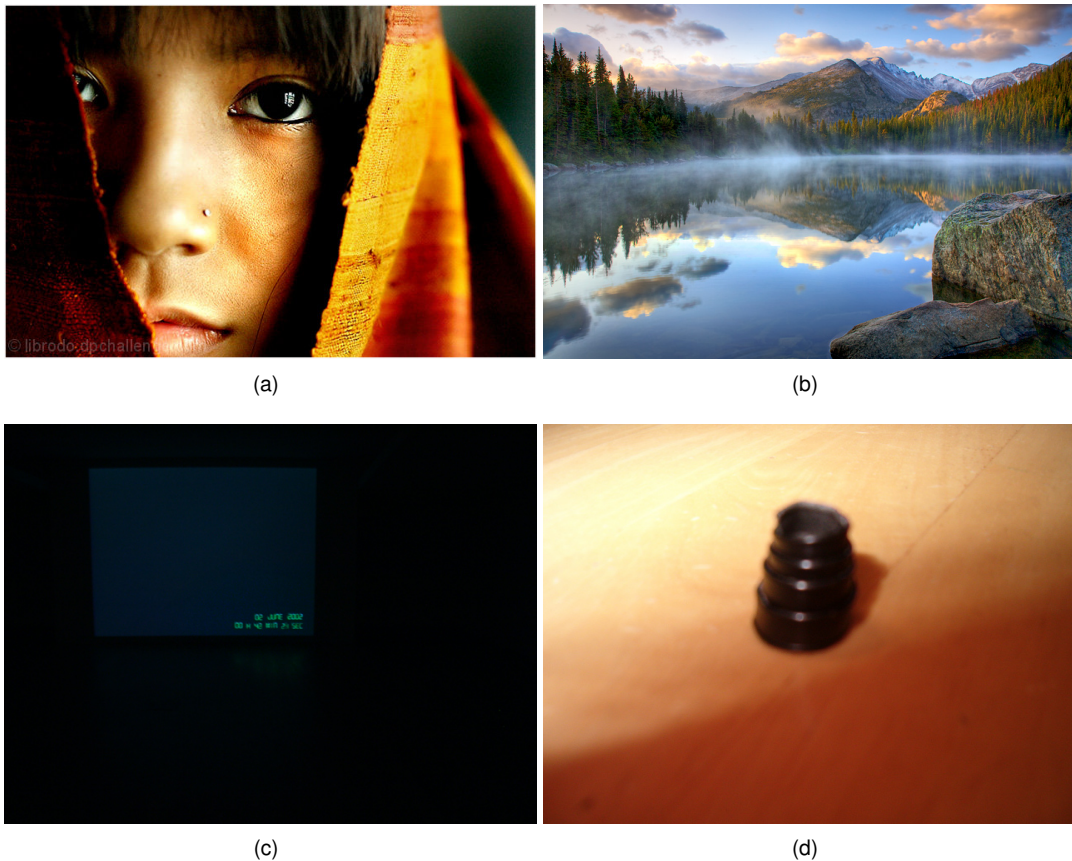


Figure 1.1: Examples of aesthetically good and bad images

Mind that under the definition of image aesthetic quality assessment, even though the aesthetic feeling of a subject can be influenced by his personal background, current interest, attention, emotion and other subjective factors, it is different from the study of emotion, interesting,

attractiveness, memorability, quality assessment, etc. They may have intersections in the biological mechanism and applications. Some researchers study images about human face's beauty or attractiveness in computer vision, but it is usually not compared with the aesthetic quality assessment we are mentioning here.

We also need to justify that the images we are concerning on in this thesis focus on part of the 2D images, especially photos. Artificial images like painting, drawing, engraving, design, film scene, etc. are not included, as the value of art pieces and designs are valued with many factors that normal photos may not have, like the artist, art style, historic value, fashion trend, etc.

There are different ways to classify photos. According to Alise Tifentale and Lev Manovich's work [104], there are many photo classes, e.g. photojournalism, commercial photography, family photography, competitive photography, vernacular photography, celebrity selfies, amateur photography, personal photography, etc. Most image data resources that image aesthetic quality assessment researchers are recently focusing on are competitive photography, family photography or vernacular photos. They explained competitive photography as the photos aim at a group of audience and peer photographers who share somehow similar mind, which follow the master practices of photographic technique, aesthetics and creativity. They are shown to the public internationally and aiming at recognition and prize or in another form, like the likeability on online social media (e.g. "likes" on Instagram, scores in DPChallenge), instead of directly financial benefits. Family photography relates to the photos focusing on the family events, relatives and friends, such as family snapshots, family photo albums. They believe that the people in the photo takes a more important part than in the competitive photography. Vernacular photos is similar to contemporary amateur photos online, they imply on lacking of technical skills and having a substandard quality of work. However, current datasets are not able to fully distinguish amateur and vernacular photos from professional photos unless they collected the photographers' information. We can see the details of existing datasets in Chapter 2. In this thesis, we are covering images classes that current popular datasets for image overall aesthetics can reach, which are a mixture of competition photography, amateur photography and vernacular photography.

The researchers improve image aesthetic assessment based on results and experience from different areas, like neuroaesthetics [12, 40], photography computation [45, 26], psychology of aesthetics, etc.

Neuroaesthetics studies the biological mechanism of aesthetic assessment. As the assessment to stimulus relate to human brain, where the visual attributes can activate the visual cortex, they chose convolutional neural network to extract visual feature. Different attributes, or

elements like lines and colours, can stimulate different parts of the brain, so the researchers can study relationship between attributes and global aesthetic assessment.

Often, researchers use tools in photography computation, which is with digital image acquisition and processing techniques. What is more, as the artists concluded some features that can move the viewers perceptually, modern photographers learned from their experience and summarised to some well-established photographic rules to make their photos aesthetically well [45]. By introducing the rules, researchers can design features and algorithms.

Psychology of aesthetics, also called experimental aesthetics, suggested different ways of getting the aesthetic values. So researchers borrowed them to get labels in image quality assessment, e.g. let a subject watch two images and choose one with respect to their aesthetically pleasing.

In fact, the aesthetic value of an image is usually directly given by a rating of aesthetics degree [45] on a scale of integers, or a binary judgement like good or bad. Thus, the aesthetic quality assessment is typically cast as a classification problem or a regression problem [26]. For an image, it can have a numerical level of the aesthetic value from one user or the average from a group of users; it can also get a group of aesthetic assessment results and form a histogram.

By using proper data and tools, algorithms are able to evaluate image aesthetic quality automatically and well. However, there still exist many unsolved problems. In this thesis, we are focusing on the problems about subjectivity and explanation of image aesthetics.

1.2 Tackled Problems of the Thesis

In this section, we are going to introduce the problems we are going to work on in this thesis.

1.2.1 Subjectivity in Aesthetic Image Assessment

Most of existing aesthetic quality prediction approaches assume that aesthetic quality can be represented by a single value, e.g., the mean aesthetic score or the aesthetic class (good/bad). However, this assumption does not take into account the intrinsic *subjectivity* of aesthetic assessment, which may be influenced by personal background, interests, mood, etc. Indeed, experimental psychology studies show that, while beauty is conveyed by objective visual clues, the resulting aesthetic appraisal is subjective and depends on how the visual clues are processed by higher-level cognitive areas in the brain [82].

As a result, summarizing aesthetic quality with a single value is not in general sufficient to capture the subjectivity of aesthetic perception. We define the subjectivity in this paper as the *degree of consensus* about the aesthetic value of a picture when the picture is judged by

a panel of human raters. The top two rows of Figure 1.2 illustrate this with an example: two images from the AVA dataset have a similar average aesthetic score, but a different degree of subjectivity. In the image in Figure 1.2(a), it is evident that humans tend to agree more on the aesthetic quality of the image, while the judgments are more dispersed for the image in Figure 1.2(b). Intuitively, being able to predict aesthetic subjectivity can provide valuable information in order to determine to which extent aesthetic predictions can be trusted. This in turn could be beneficial in applications such as enhancement or retrieval, in order to obtain more reliable and accurate results.

Recent work has tackled aesthetic subjectivity by predicting the *distribution* of subjective scores of an image [43, 76, 101, 44]. Specifically, these methods leverage the availability of ground-truth aesthetic score distributions obtained by a large number of human annotations, offered by large-scale datasets such as AVA [77], and employ different loss functions to measure the distance between probability distributions. However, aesthetic subjectivity is described only implicitly by the score distribution. Instead, we are interested in quantifying and predicting this subjectivity *explicitly* through a scalar value that summarizes the score distribution by describing the raters' consensus, and which could be used by automatic image analysis algorithms.

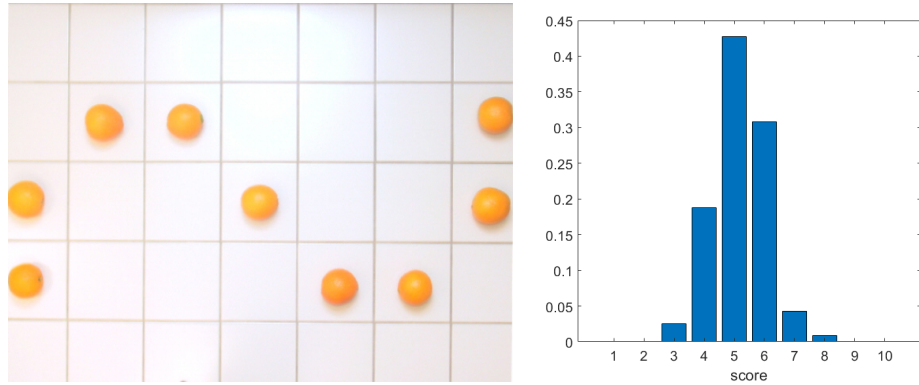
Therefore, in this thesis we want to analyze the measures of subjectivity based on the distribution of the scores for better understanding the aesthetics.

1.2.2 Explanation of Computational Aesthetics

What makes an image aesthetically beautiful?

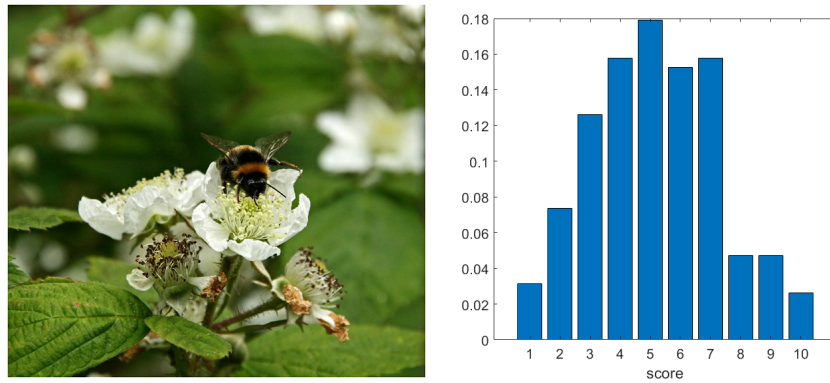
It is challenging to draw conclusions on how each aspect affects the aesthetic score. A typical approach is to define features involving different attributes, which are then integrated into classification and regression models to match the scores given by observers [26]. For example, attributes such as colorfulness, tone, clarity, depth, and sharpness are computed in [2]. With the growing interest for deep learning methods, several works have tried to add high-level attributes which cannot be well explained by hand-crafted features. In [68, 50, 31], the authors have verified how high-level attributes, like style and semantic, affect aesthetic scores.

Besides the mentioned attributes, genres of the photos is believed to affect the aesthetics. First of all, photographers are believed to shoot differently from the un-trained [105]. From content category's aspect, techniques are taught differently in photographic books for categories like portrait, scenery, animals, etc. As an example of scenery photography, choices of vivid colors, tonality, composition rules like the rule-of-third and diagonal lines, focus, lighting, etc., are emphasised as common rules in producing a high aesthetic level photograph. From the shooting equipment's aspect, single-lens reflex camera, mirror-less camera, digital camera, mobile



Mean	SD
5.174	2.104

(a)



Mean	SD
5.180	0.902

(b)

Figure 1.2: Example of aesthetic subjectivity for two images of the AVA dataset. The two images, displayed in the top-left panels, have similar mean score but different distribution of aesthetic judgments given by human raters, shown in the histograms on the top-right panels. The tables report the mean score and standard deviation of the votes.

phones and even some specific type of film have various of features and visual quality. From the application aspect, images like HDR images, thumbnails and posters need to implement different skills to motivate or immerse the viewers.

Aside from these factors of an image, since existing aesthetic values are collected from human, whether the characteristics of subjects, like colour visual condition and personal background, affect the assessed aesthetic values or not is not fully studied.

Over all, what is making an image aesthetically well is not studied enough.

1.2.3 Human-annotated Data Collection

Since aesthetic quality assessment is a subjective task involved with human, how to collect human's data reliably and use them has been considered by many researchers. The task and attributes definitions are often ambiguous and subjective, which limits the reliability of the collecting methods. One of the recurring issues that hinder the progress of research in aesthetics is the lack of properly labelled data.

There are two ways of collecting data. Inferring aesthetic information indirectly from human raters has advantage in collecting lots of data in a short time, like crawling from websites. Yet, it is difficult to disentangle and quantify the factors producing the aesthetic appraisal. Thus, collecting aesthetic annotations directly from users remains the most reliable way to provide accurate ground truth labels to predict aesthetic quality. However, this way requires resource in time and finance, and the experiments should be designed properly to avoid too much noisy data.

Choosing images improperly can influence the aesthetic judgement of subjects. Barthes proposed that a press photo is a message and is formed by the source of emission, a channel of transmission and a point of reception [4]. In image aesthetic quality assessment data, we want to collect the evaluation of aesthetic message that transmitted from the photo-taker by his photo to the viewer. Previous research limited derived the influence of connoted messages (similar with *studium*, it is related to cultural, linguistic, and political interpretation of a photograph [3]) when they choose images and subjects in data collection. For example, commercial photos like advertising photos with clear and distinct meaning [3] and "mask" photos with social critique are mixed in the image collection in different extent. This may because photography consider more than aesthetic level, and many platforms are open to the public to collect photos, so the datasets who crawled data from them contain these images.

As algorithms are tested on datasets, the lack of well-labelled data is negatively impacting progress towards not only prediction, but also understanding the importance of attributes in image aesthetics. We need new datasets with a reasonable variety and quantity of annotated

images.

1.3 Objectives of the Thesis

The objectives of this thesis can be summarized as follows:

- Measure aesthetic subjectivity.

We define subjectivity as the degree of consensus of human raters, and uncertainty as the individual confidence in a rating. Since there was no study on the subjectivity of aesthetic score of images, we need to define and measure subjectivity. With quantified subjectivity, we can know if an image is definitely good or bad to a group of raters. After having the measurements, We can evaluate the prediction accuracy of the proposed subjectivity measures, and compare these results with directly learning the subjectivity scores. We want to see if predicting subjectivity directly is better than predicting distributions and then compute the subjectivity measure. Also, we can test if subjectivity measures can help improving aesthetics prediction.

- Build a controlled explainable aesthetic dataset.

We propose a new dataset that simultaneously contains subjective labels for aesthetic attributes ranging from low-level visual factors (e.g., light, color, exposure) to higher-level semantic preference. In addition, we record the importance of each attribute while collecting votes, by explicitly asking observers to indicate which factors influenced their aesthetic judgment for a given image. We also ask for the uncertainty of subjects' votes while voting. This is meaningful to evaluate the average aesthetic score's reliability. At the same time, we combine crowd-sourcing with the best practices from quality assessment recommendations such as subject training and a clear definition of the attributes to select test stimuli and guarantee the quality of the collected data. Personal background and demographic information are also collected. Finally, user voting time is recorded in order to identify outliers and clean the data.

- Investigate of the relationship between uncertainty and subjectivity.

We are interested in the uncertainty data, which is orthogonal to subjectivity in that it describes the personal degree of confidence on the ability to assess aesthetically an image. Since there is little study about group disagreement and difficulty in judging image aesthetics, we want to study if they have relationship.

- Figure out the factors leading to a certain aesthetic quality.

We employ our dataset to learn how to disentangle aesthetic attributes from the overall aesthetic score. We want to know among the studied attributes (light and colour, composition and depth, image quality, semantics), which attribute is the most important in the overall aesthetic quality of images, and if different content categories display a different relative importance of the attributes.

1.4 Structure of the Thesis

The rest of the thesis is organized as follows.

- **Chapter 2 Related Work.** We review the relevant work to this thesis, summarized to four parts: (1) State-of-the-art of image aesthetic quality prediction. (2) Predicting aesthetic subjectivity. (3) Existing datasets and collection methods. (4) Work towards explanation of aesthetics.
- **Chapter 3 Modeling and Predicting Subjectivity in Aesthetics.** We present several subjectivity descriptors and evaluate different prediction schemes for these measures, and compared predicting subjectivity directly with predicting distributions and then compute the subjectivity. We use the subjectivity measurements to improve aesthetic score prediction.
- **Chapter 4 EVA: an Explainable Visual Aesthetics dataset.** We present the Explainable Visual Aesthetic Dataset, including the choice of stimuli, the design of the questions; the methodology; the data collection (recruitment of participants, removal of outliers), etc.
- **Chapter 5 Aesthetics and Attributes.** We analyse the collected data to find out the influencing attribute to aesthetically good images and the relation between attributes and difficulty. We also tested whether user's characteristics influence their average aesthetic assessment.
- **Chapter 6 Conclusion and Perspective.** We conclude the thesis and discuss perspective in this field.

Chapter 2

Related Work

General aesthetics prediction deals with predicting image aesthetics for any kind of image content, in contrast to task-specific aesthetics where the class or object of the picture is known, e.g., images of faces [6]. In this thesis we focus on the general image aesthetics problem regardless of artistic value and contextual information.

Since the subjectivity involves deep learning regression, and the collected dataset can be used for aesthetics prediction, we first give an introduction in deep neural networks and image aesthetic assessment in Section 2.1. Then, we review the related work to subjectivity in Section 2.2. In Section 2.3, we discuss about explainable image aesthetics. In Section 2.4, we go through existed datasets in image aesthetic quality assessment.

2.1 Image Aesthetic Quality Prediction

In this section, we are going to talk about the state-of-the-art and basic pipeline of image aesthetic quality assessment, and review a few popular convolutional neural networks commonly used in regression task.

2.1.1 Aesthetic Assessment Pipeline and Deep Neural Networks

After getting a new image without an aesthetic value representation, the first step is to learn a feature from it. This feature can be attributes that can construct the aesthetic values, or a description of the whole image. This representation can be designed or learned. Then, there is usually a model to build the aesthetic value, which involves a decision procedure. There is usually a criteria to evaluate the performance of the learning model, which includes classification accuracy, precision-and-recall curve, Euclidean distance and residual sum-of-squares error

between the predicted value and ground-truth value, correlation ranking, ROC curve, mean average precision, mean square error[26], etc. Subjective evaluation is also used as criteria, but is more commonly seen in the procedure of collecting the ground-truth ratings in building the datasets.

Initial works used handcrafted features to represent images' visual attributes[45] or a quality of the image itself to model the aesthetic values. They are based on handcrafted signal filter design. At this time, training these features and learning a classifier is an important part of this field, and choosing the kind of learning algorithm depends on the data. Condition of dataset and applications decide which type researchers should use. Three types of learning strategies are used in this field: supervised learning, semi-supervised learning and unsupervised learning.

As the rapid development of deep learning, Deep Neural Network (DNN) framework has become the most popular tool in academic. Learning aesthetically sensitive deep features and aesthetic task classifiers directly from the images and aesthetic data is the most common way nowadays.

Neural networks were discovered in 1940s in the neurophysiology [74]. In the past two decades, hardware development and big datasets allowed the training of deep neural networks (DNN) to be more feasible [88]. The paper published by Krizhevsky et al. in 2012 [56] proposed a deep convolutional neural network for image classification that reached high performance with the help of the ImageNet dataset [25]. Ever since then, the deep neural network was adopted to many fields, such as computer vision and image processing [96, 100], natural language processing [14, 79], robotics [62], big data [32], etc. Plenty of network architectures were invented to suit complex tasks in real life. Their complexity increases, the networks gets deeper and more powerful, and the performance increases fast.

Different network architectures are widely used, while Convolutional Neural Network (CNN) is the major class. CNNs follow similar design patterns: they can consist of several convolution layers, activation layers, pooling layers, normalization layers and fully connected layers. These layers can form a single column network, multi-column network, single task network, multi-task network, etc.

A brief introduction to three networks are given, as they will be mentioned in the following chapters.

AlexNet [56] is the earliest successful deep neural network model in ImageNet classification task that made people realize the capacity of neural network. Its structure is shown in Figure 2.1 [56]. The network requires a $224 \times 224 \times 3$ input image, and goes through 5 convolution layers. Then, after using Relu activation to add non-linearity of the values, and a max pooling layer is used to reduce the feature's dimension. After that, it is fed into fully connected layers.

A 1000-way softmax is implemented on the output at the end. This network reached a 10% improvement in classification task at that time.

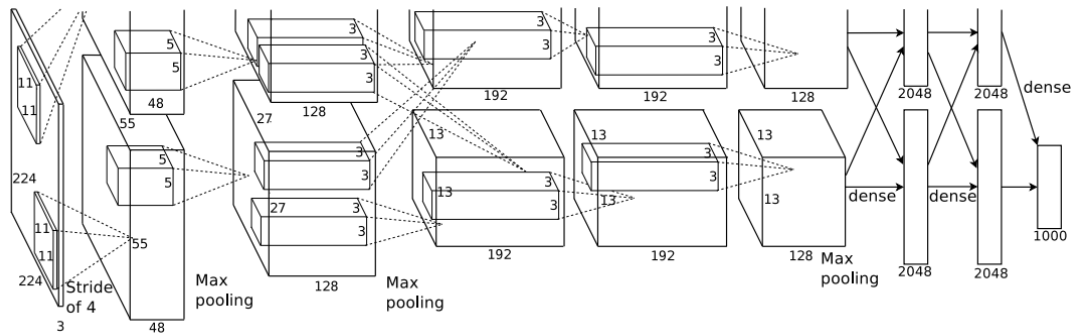


Figure 2.1: Network architecture of AlexNet [56]

VGG [96] was proposed later by Simonyan et al. and reached a higher performance than AlexNet on the ImageNet classification task. It has many forms composed of different number of layers, including VGG-11, -13, -16 and -19. The most commonly used one in image aesthetics assessment is the VGG16, due to its relatively good efficiency and performance comparing to previous models, and the simplicity in its family. The structure is shown in Figure 2.2. It requires the input image to be $224 \times 224 \times 3$. Comparing to AlexNet, VGG16 changed the convolution kernels, from 11×11 and 5×5 to 3×3 .



Figure 2.2: Network architecture of VGG-16 [96]

ResNet [37] architecture family is proposed by He et al from Microsoft Research Asia (MSRA). It consists of ResNet-18, -34, -50, -101, -152 and -1202. They have residual blocks to solve the problem of vanishing gradient in ultra-deep networks. They also requires the input image to be $224 \times 224 \times 3$. The ResNet-34 structure is shown in Figure 2.3. The most frequently used structures are ResNet-34, -50 and -101. Comparing to VGG-16, ResNet-34 takes less computing resource in aesthetic assessment architecture in large amount of data, and has a similar performance [5]. ResNet-101 and 1202 are good in performance but take large memory and longer training time, so they are usually used for improving the accuracy. ResNet-34 and -50 are more often used by other researchers in exploring the algorithm.

The input of the Convolutional Neural Networks(CNN) in the off-line training session is the

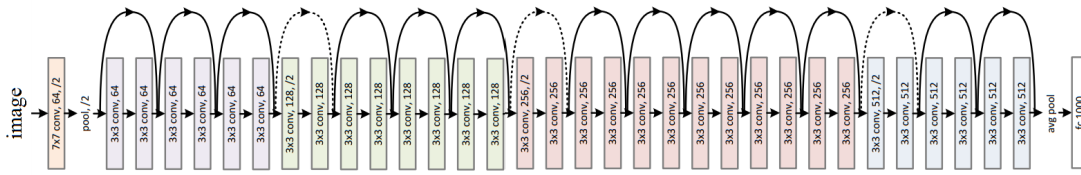


Figure 2.3: Network architecture of ResNet-34 [37]

images. They usually follow a fixed size to guarantee that the output is in a constant scale. The network can be initialized randomly or be pre-trained with other tasks. For example, PyTorch and other popular frameworks provide pre-trained models on ImageNet. After the image passes through the networks, we can get a loss value from the designed loss function by comparing the output with the ground-truth. By having the loss value, the network can automatically adjust the training direction and amplitude with back propagation mechanism. For the on-line testing session, images without labels are put into the network where parameters are fixed, and then we can get a predicted scalar or a vector.

Thank to its powerful, generalise and robust feature learning ability on large amount of images, CNNs perform well in aesthetics-involving tasks, aesthetics classification task, and predicting aesthetic values as a regression.

2.1.2 Background of Predicting Aesthetic quality from Images

Aesthetic quality prediction has received an increasing attention in the past few years in the multimedia community. Some methods focus on predicting overall aesthetic value applications, while some focus on personal aesthetics prediction. The aim is to increase a model's performance in aesthetic datasets.

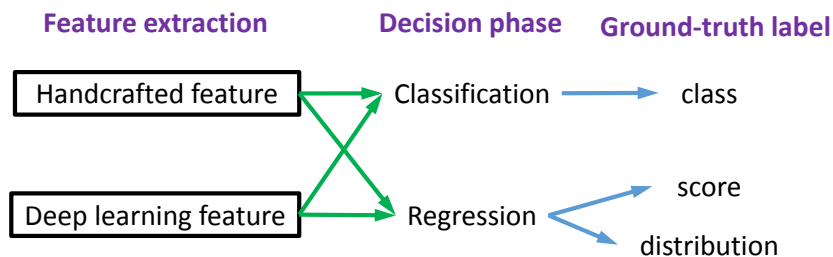


Figure 2.4: Scope of image aesthetic quality prediction

Figure 2.4 shows the scope of image aesthetic quality prediction. In the step to extract a feature from the images, there are usually handcrafted features and deep learning features. For the decision phase, the assessment is usually considered as a classification or a regression.

Classification usually has aesthetic classes as ground-truth labels, and regression can have an aesthetic score or a distribution of a group of votes to be the ground-truth label.

Besides, some researchers found aesthetic assessment task has some specific problems. For example, changing image size to suit the network input requirement may cause a change of its aesthetic value. Thus, they introduced methods like attention mechanism, pooling layers, patches, etc., to make the algorithm suitable for more sophisticated aesthetic task.

General Aesthetics Classification

Many methods tried to find a way to distinguish photographers' photos from unprofessional photos, assuming that photographers' photos have a higher aesthetic quality than normal people's. Ke et al. [53] proposed algorithm based on their observation on traits of photographer's photos and snapshots. They found simplicity of the photos, realism and the use of special photographic techniques can distinguish them. Thus they compared spatial edge map, colour, hue, blur, contrast and brightness on images collected from *DPChallenge.com*. In [117], authors introduced the function of saliency of image in aesthetic quality binary classification.

These handcrafted features involved great work of defining new features and improving the mathematical models, and it highly relies on the researcher's professional understanding of the targeting images, filters and experience of photography. At the same time, from the way of filtering the useful attributes, we can see that the workload is heavy. In different datasets, the useful attributes may change. Moreover, as the image acquisition technique develops, the aesthetic quality of non-photographers' photos are getting better, which limits the human observing and manually design of new discriminators.

Recent methods consider image aesthetics are labeled in a classification way, not by the photographer's levels. They collect people's opinion on image aesthetics, and set low and high thresholds for image's mean opinion score. Often, they neglect the images within medium score scale. Inside each class, images are considered as the same aesthetic level, so that they need not to have a higher granularity.

There are works trying to merge traditional hand-crafted features to neural networks. Authors in [73] figured out that the generic descriptors of the image aesthetics is being encoded explicitly by content based descriptors, which later inspired Dong et al. [29] to use the generic features in the penultimate layer of AlexNet with spatial pyramid pooling to predict the binary class. In [107], Wan and Tian continued their work in combining traditional machine learning or signal processing tools into deep learning architectures, as they first use traditional attributes to get binary classification labels, then use three classification networks to extract the attribute descriptors. Then, they combined them as three columns, which are in parallel of another column

Table 2.1: Information utilised in general image aesthetics prediction

Information resource	Information
Observation	photographer's visual and photographic difference
Hand-crafted visual attributes	photography rules
Deep features	pre-trained attribute features local features global features
Collected labels	visual and photographic ratings semantics/object scene style emotion
Text comments	visual attribute words photographic attribute words high-level attribute key words

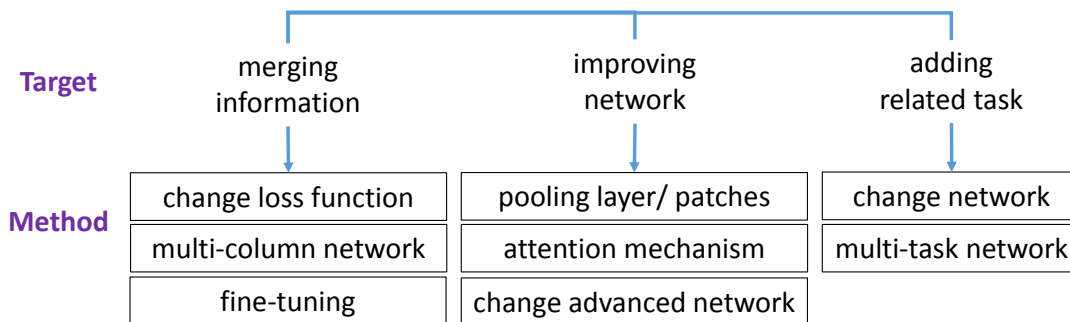


Figure 2.5: Scope of improving aesthetic value prediction methods

that represents the learning of the rest of global features, to compose a multi-column network. However, the features rely on manual design for classification pre-training, which may introduce a bias. Kucer et al. [57] fused handcrafted features with deep learning features to improve accuracy.

RAPID model [67] is literally considered as the first attempt to using Convolutional Neural Network (CNN) to train aesthetic feature and classification model. It used AlexNet [56] and changed the last fully-connected layer to output two classes. Later, it is improved to DCNN [69], as a combination of two network columns to catch the global feature and local features at the same time. Furthermore, authors boosted the accuracy in binary classification by introducing a third style/semantic information column, and named it SDCNN.

After these first successes of deep neural network, related work boosted. Most of the work focused on changing network structure or adding information in training. Multi-scene deep learning model (MSDLM) [112] was first trained to classify different scenes, then merged to predict the aesthetic score. Kong et al. [55] used a unified architecture based on AlexNet to

consider aesthetics, useful attributes and photo content at the same time, and they modified the sampling strategy that mixes the inter-rater and intra-rater image pair comparison to get a ranking loss. Kao et al. [50] used multi-task structure to predict aesthetic binary class and semantic class in their multi-task convolutional neural network (MTCNN). Lee et al.[60] studied an encoding method for high-dimensional CNN features. Zhang et al. [120] fused last layers of two network trained on different datasets in their hierarchical features fusion aesthetic assessment (HFFAA) model to obtain the functions of object-aware and scene-aware at the same time. Gao et al. [31] proposed training binary aesthetic classification model first, then using a support vector machine to predict a style classifier, and using multi-task learning to learn aesthetics specifically with style.

Researchers also tried to import extra text information into visual aesthetic methods to discover about aesthetics, since human text words is a reflection of people's opinion. Zhou et al. [123] proposed jointly learning aesthetics with images' text comments. Yu et al. [119] combined emotion prediction with aesthetics.

Personalise-aimed Task

Differently from general aesthetics, personal aesthetic prediction study is done to fit the diversity of aesthetics. Park et al. [80] consider personal taste in addition to general aesthetic score, by adapting a model to match specific user preferences obtained from user interactions. They implemented this joint learning aesthetic value regression model and user's ranking preference model on a small dataset. Authors in [84] learned from Inception-BN network to build a residual-based personalized model to predict aesthetic scores. They utilise a support vector regressor to predict the residual score, and use aesthetic and content attributes to predict a generic aesthetic value. Wang et al. [108] added text information to do a multi-modal personal aesthetic prediction. Lee et al. [61] predict aesthetic binary class, score and personalized score at the same time with a single input image. Wang et al. [113] introduced meta-learning framework to solve the lack of labelling in this direction. For personalised preference, besides each one's own conditions, people are also influenced by their surrounding social networks. Cui et al. [22] catch users' favouring behaviour on social media platforms as a reflection of their hidden cognition preference to images.

Study of the Aesthetic Networks

A few works focused on the technical problems caused by deep neural network in the aesthetic field. In tasks like recognition, input images are resized, cropped or patched to a fixed square size, so that the output features can share the same dimension number. However, in aesthet-

ics, it is commonly believed that these pre-processing actions have a possibility to change the aesthetic value to a human. The network may not learn the aesthetic difference between the processed image and an original image as human does.

To solve this problem, there are two popular trends. One is utilising patches or boxes to get part of the image details, then combine with global features. Another one is by using spatial pooling layer to implement a dimensional reduction to connect fully connected layers. In [72], Mai et al. used an adaptive spatial pooling layer between the VGG-based convolutional layers and the fully connected layers, so that the input images can keep its original size and aspect ratio. Wang et al. [111] used similar pooling layer to deal with arbitrary size of image, and they introduced attention-aware cropping function in the architecture to crop a beautiful photo from the given image to reach the need of enhancement. Semantic-Aware hybrid NEtwork (SANE) [21] and View Finding Network (VFN) [13] also use pooling layer to deal with the input size. Adaptive Layout-Aware Multi-Patch Convolutional Neural Network (A-Lamp CNN) [71] is also for studying accepting arbitrary input image size and ratio. It has two main columns, one for acquiring features of salient objects in patches in an image, another one for learning the global topology and spatial feature of the image.

Attention mechanism based ideas appeared in trying to effectively select information in learning as human pays attention to an image. Besides work like [71], Kao et al. [49] proposed predicting scene, object and texture category simultaneously with predicting aesthetics binary class, instead of three separate columns of network for classification. Brain-Inspired Deep Networks (BDN) [115] improved RAPID and DCNN by pretraining attributes first, and then the researchers improved the work and proposed Deep Chatterjee's Machine (DCM) [116], which is inspired by the Chatterjee's visual model in neuroscience. Semi-supervised learning method in [66] involved human gaze shifting path (GSP), believing that is the way human get aesthetic information in noisy data, and did an experiment with an eye-tracker. [92] proposed an attention-based multi-patches aggregation mechanism that strengthens incorrectly predicted patches, and [110] studied an attractive region search algorithm. Zhang et al. [121] were inspired by human perception of fine-grained details with a mixture of foveal vision and peripheral vision. They also improved the fusion of global and local information based on attention mechanism [122].

At the same time, the invention of new architectures, methods and tools not only improved the performance, but also gave more possibility in this field. Deng et al. [27] introduced EnhancedGAN to generate an enhanced image. Weakly supervised is paid attention in aesthetically auto-cropping [63, 64], based on reinforcement learning. [86] tried to use object recognition result and some context tags to train aesthetics. [114] added a Long Short-Term Mem-

ory(LSTM) part that can generate text from the aesthetic visual attributes. Self-supervised learning is introduced for solving high human labelling cost, but the study is limited in the image aesthetics [93] and the importance of human labelling is not replaceable yet. Meta-learning is also a direction for solving lacking labels by learning information from similar task and adapting to aesthetic task [113].

Aesthetic Regression

Most of the above mentioned works are increasing the classification accuracy, but regression is also getting popular, as it suits more fine-grained prediction requiring applications.

Kao et al. [48] is the first work in predicting the aesthetic score as a regression task involving deep neural network. They input cropped images to their network, and make the output to be a single numerical result. They use mean residual sum of squares error as an evaluation of the network, which is inherited by most of later work.

Bin Jin et al. [43] proposed using inversely proportional weights to emphasise the samples that are not well learned by the network due to the occurrence times in training. The weights are computed with the occurrence times of the mean score in the dataset, as in Function 2.1. They divided mean scores to bins by value, and b_i is the occurrence time of the i th bin's mean scores. They first used weighted mean square error as a loss function for predicting the aesthetic scores, which means the $dist(i) = (y_i - \hat{y}_i)^2$ in function 2.1. y_i and \hat{y}_i are the predicted and ground-truth mean score of i th image. Then, they used the weight in predicting the distribution of aesthetic scores by a weighted mean χ^2 error loss, where $dist(i) = \chi^2(h_i, \hat{h}_i)$, and h_i is the predicted histogram, \hat{h}_i is the ground-truth normalised histogram. It can account for the non-uniform distribution of the scores in the AVA dataset. Their experiment used a modified VGG-16[96] network where fully connected layers are changed to get 1 or 10 dimensional output for score and distribution respectively. In the end, they implemented the algorithm in image cropping application.

$$w_i = \frac{\sum_{i=1}^B b_i}{b_i} \quad (2.1)$$

$$WMSE = \frac{1}{\sum_{i=1}^N w_i} \sum_{i=1}^N w_i * dist(i)$$

Later, Murray et al. [76] employ the Huber loss and spatial pyramid pooling [36] to predict distributions. Huber loss is known to be less sensitive to outliers. Their network is modified based on ResNet-101.

In the NIMA system [101], the loss function consists of the squared Earth Mover's Distance (EMD), as shown in Function 2.2. CDF_{pk} is the ground-truth cumulative distribution function, and $CDF_{\hat{p}k}$ is the one computed from predicted histogram. This loss function can consider the

intrinsic orderness of classes. They also tried their loss function on three different networks, including VGG-16, and the result as regression and binary classification accuracy showed an improvement.

$$EMD(p_i, \hat{p}) = \frac{1}{N} \sum_{i=1}^N |CDF_{p_i} - CDF_{\hat{p}}|^2 \quad (2.2)$$

The authors of [44] propose to use the cumulative Jensen-Shannon Divergence based CNN (CJS-CNN) as loss function for predicting aesthetic distribution. The cumulative Jensen-Shannon Divergence [78] allows non-parametric computation and it is more robust facing the curse of dimensionality. When it is used in the loss function, it can detect the change of distributions efficiently. The authors present an extended version of this loss using a function of the kurtosis of ground-truth score distribution to weigh CJS (RS-CJS). Kurtosis is used as a proxy to aesthetic "reliability", and used to penalize more those images whose distributions are considered unreliable, as proposed as Function 2.3 in their paper. y is the ground-truth histogram and \hat{y} is the predicted histogram. Th is a threshold defined by the data in the training set. $r^{kurtosis}$ is a measure of reliability. When the shape of histogram is close enough to a normal distribution, it is more reliable. They also carry out a comparison of their method with several other distribution losses, showing that RS-CJS provides the most accurate distribution estimations according to several performance criteria.

$$T(y) = \frac{1}{|kus(y) - 3|},$$

$$r^{kurtosis}(y) = \begin{cases} \frac{\ln(t(y)+1)}{\ln(t(y)+1)+1} & T(y) < Th \\ 1 & \text{otherwise} \end{cases} \quad (2.3)$$

$$l^{RS-CJS}(y, \hat{y}) = r^{kurtosis}(y) CJS(y, \hat{y})$$

2.2 Predicting Aesthetic Subjectivity

With the advantage of deep learning and the availability of large-scale datasets with thousands or hundreds of thousands pictures [77, 55], the accuracy of aesthetics prediction has been constantly improving. In this context, the problem of aesthetic assessment has been mainly formulated as predicting the average score or the aesthetic class of an image [67, 55, 69, 26]. The image aesthetic level classification [67, 69, 50] work focuses on dividing images mainly to two classes- high or low quality. This hard division usually have a threshold in labels of training data and ignore the aesthetics prediction with medium level images. Comparing to classification problems, regression considers all images on the scale, and we can know the image's aesthetic

position in a rating scale by having a scalar number, which is good for applications like auto-enhancement and recommendation. The first image aesthetic value regression task is done by Kao et al. [48]. After that, people tried to predict the average of group opinion, but they studied group diversity little.

Although there are studies on personalised aesthetics prediction to catch the difference between individuals, they can not describe the group diversity for aesthetics. We do not target personalized aesthetic prediction, but rather aim at assessing the level of consensus of a panel of humans about the general aesthetic value of a picture.

Aesthetic subjectivity, as the *degree of consensus* about the aesthetic value of a picture, has been considered recently by a few studies by predicting the distribution of human scores, rather than simply the mean opinion score.

Some works consider the problem as a regression of score distributions [43, 76, 101, 44], because they suppose that once the distribution is known, one can compute its moments, e.g., mean score and the standard deviation.

Jin et al. [43] proposed that the predicted distributions can indicate the difficulty of image's aesthetic assessment by the standard deviation, besides the mean score. Murray et al. [76] noticed that some very skewed images do not have a skewed prediction, but have a nearly symmetrical distribution with a mean of the average rating score.

NIMA system [101] consider the intrinsic orderness of classes, but we can see from Figure 2.6 that their predicted standard deviation is more gathered than the ground-truth standard deviation.

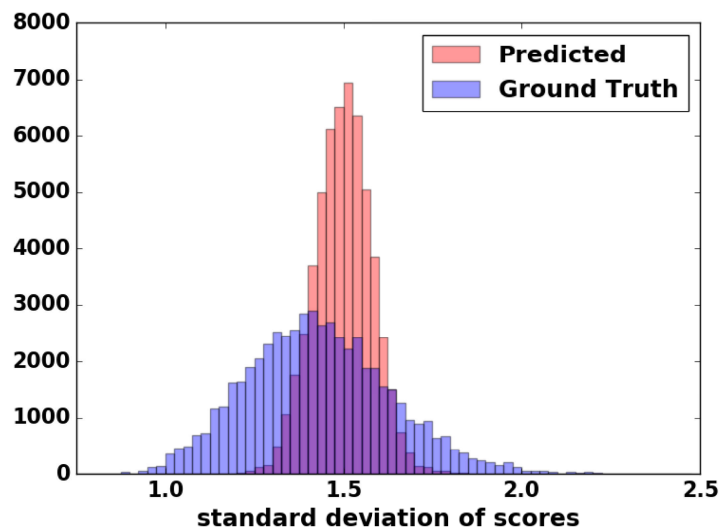


Figure 2.6: Distribution of predicted standard deviation and ground-truth's standard deviation [101].

Jin et al. [44] clearly noticed the aesthetic quality rating histogram is subjective. They

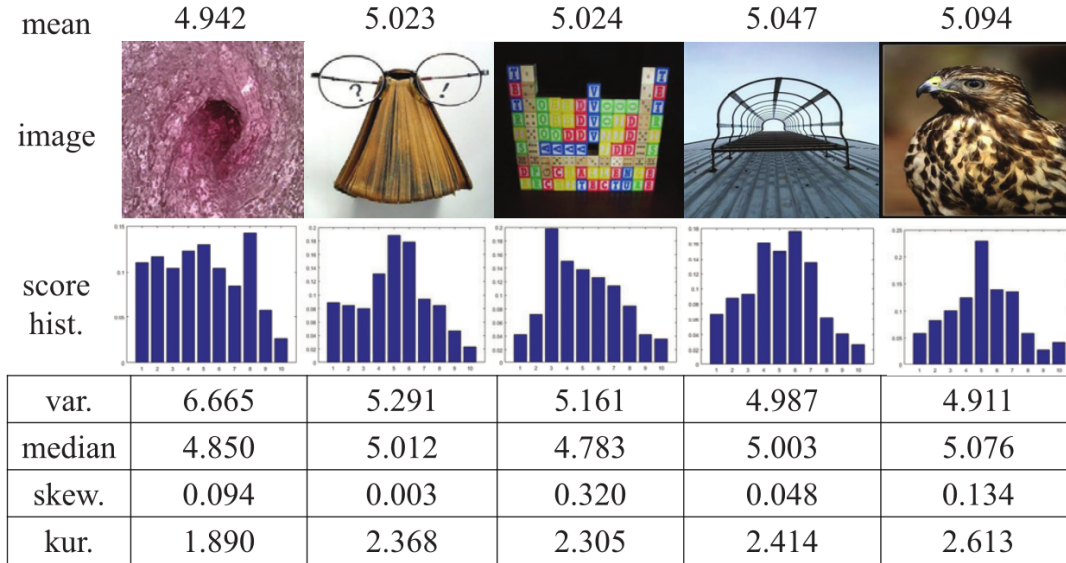


Figure 2.7: Images with similar mean scores share different histograms, variance, skewness and kurtosis [44].

used Jensen-Shannon Divergence to keep the subjectiveness. They noticed with similar mean scores, ground-truth histograms share different variance, median, skewness and kurtosis among images, as shown in Figure 2.7. They chose kurtosis to measure the reliability of a score distribution. The key functions are shown in function 2.3. Kurtosis measures how heavily the tails of a distribution differ from the tails of a normal distribution. $kus(y)$ is bigger than 3 when distribution y has a thin tail, and smaller than 3 when the distribution has a fat tail. Thus, from the function, it means that if the distribution is closed enough to normal distribution, the distribution is more reliable, and the error should be more penalised in training. However, the intrinsic disagreement is different from this reliability, and histogram can do not follow normal distribution but have a high consensus. They did not use explicit measure of subjectivity to evaluate the considered methods.

In this thesis, instead of predicting the score distributions, we propose(for the first time to our knowledge) to define explicitly subjectivity measures and directly predict them.

Besides, in [38], Mean Opinion Score (MOS) is not enough for describing the diversity of the user ratings for a subjective task in Quality of Experience (QoE) study, and they proposed that the standard deviation of the ratings (SOS) may be related to MOS in a quadratic function. Distribution prediction methods assume that after predicting the distributions, it is easy to get the subjectivity or other measurements, but whether this is valid or not is unknown. Thus, it is useful to know if getting subjectivity from predicted distribution is better than learn to predict subjectivity directly. We will also study this in the thesis.

2.3 Explanation Attempts in Images Aesthetics

Many attributes are considered affecting aesthetics, where visual attributes are the most widely studied. For example, authors in [15] studied if the image showing size and resolution can affect the perceptual beauty. Authors in [2] aim at exploring an aesthetic signature composing of visual aspects in attributes such as colorfulness, tone, clarity, depth and sharpness to summarily represent the image.

Photographic attributes are popular as well. Authors in [23] chose the color features related to photography and color psychology in HSV color space, and they computed representations of homogeneous regions to describe the objects in an image. Then, they manually designed 56 group of features from light exposure, colorfulness, saturation, hue, rule of thirds, human familiarity, texture, image size, aspect ration, region composition, depth of field and shape convexity. After that, they figured out top 15 features after a 5-fold cross-validation that can effectively distinguish aesthetically pleasing and unpleasing images with the images from *Photo.net*. [107] tried color, layout and clarity factors. Dong et al.[28] proposed color, foreground and background separation, sharpness, depth-of-field and image size can affect aesthetic quality regardless of different datasets.

Several works have tried to add high-level attributes which cannot be well explained by hand-crafted features. In [68], the authors have verified that high-level attributes, like style and semantics, can affect aesthetic scores.

Besides observing, several ways are used to verify these attributes are affecting image aesthetics.

Early attempts tested if different features can distinguish professional photos from photos by ordinary people, supposing photographers can shoot aesthetically better photos[53, 117]. However, due to the robustness and explosion of image quantity in use, the conclusions nowadays need to be renewed, and they did not quantitatively describe the relationship between attributes and the aesthetic score.

Then, researchers build datasets that not only consist of general aesthetic values, but also attribute labels. The popular benchmark dataset AVA [77] collected voting scores “in the wild” and related image data from DPChallenge, where photography amateur competitions are held online. Subsequently, the dataset has been expanded [68], to meet the needs of deep learning models, but AVA remains the most popular dataset in aesthetic studies. Several datasets have continued to crawl data based on AVA to augment information from the users and photographers. However, they have many disadvantages in the labelling and scale, which we will discuss in section 2.4.

With the datasets, a typical attempt is to define features involving different attributes, which are then integrated into classification and regression models to match the general scores given by observers [26]. Also, correlation coefficients can be computed to show the relationship between the attributes and subjective score. What is more, if the attribute is making the predicted score closer to the ground-truth, it is usually believed to have an affect on the aesthetic value of an image.

For example, in [2], they selected the top 15 among the 56 designed influencing features, and the accuracy of combined features for classifying high and low aesthetic quality can reach to around 70%, and a single feature's classification accuracy can be over 50%. Authors in [15] collected users' ratings in crowd-sourcing, and then selected handcrafted features to build a model for predicting aesthetic score and ranking. Then, they found image size and resolution is affecting the perceptual beauty depending on the image content. They concluded that content needs to be considered in future aesthetic assessment. [65] provided a dataset by doing subjective user study in a controlled lab condition, but the aesthetic attributes are computed from the images, not by collecting human opinions. Their result shows no aesthetics features that they studied is significantly correlated with the continuous image aesthetic scores, and image aesthetics with different contents have small relationship with different types of features.

With the development of the deep neural networks, this method is getting easier and easier because of the outstanding ability of extracting features. In [107], authors use a multi-column neural network to first train color, layout and clarity factors, then combine them together with a column for undefined attributes, to imitate the general aesthetic values. Here, the training of the visual factors are based on the labels classified by K-means clustering the computed values. This way of labelling can introduce bias from the beginning of training, as the attribute values are manually designed and the classification does not reach 100% accuracy.

These approaches reached good performance and are good for practice, but they cannot provide an interpretable explanation of image aesthetics, nor comparing attributes' importance to aesthetics.

A few works of explaining aesthetics were done. Aydin et al. [2] illustrated Figure 2.8, where people can directly see the importance of each attribute and compare the attribute difference in an edited image with the original one, but they only considered low level features and the attributes are manually designed. Chang et al. [11] tried to generate comments automatically by having an image. The authors in [114, 123] extracted the aesthetic attributes importance based on text comments. However, the attributes definitions are ambiguous and may differ for each user, which limits the reliability of the method. In AADB [55], the subjects have been asked to vote for 11 attributes in three scales directly at the same time by the same users across

multiple images. Still, the study was only conducted with 5 or 6 observers per image, recruited using Amazon Mechanical Turk(AMT).

These methods have the advantage to provide an interpretable explanation of image aesthetics, but fail in capturing accurately complex aesthetic phenomena. As a result, these approaches tend to perform poorly when tested in real-world conditions with a wide content variety. Generally speaking, the lack of well-labeled data is negatively impacting progress towards understanding the importance of attributes in image aesthetics.

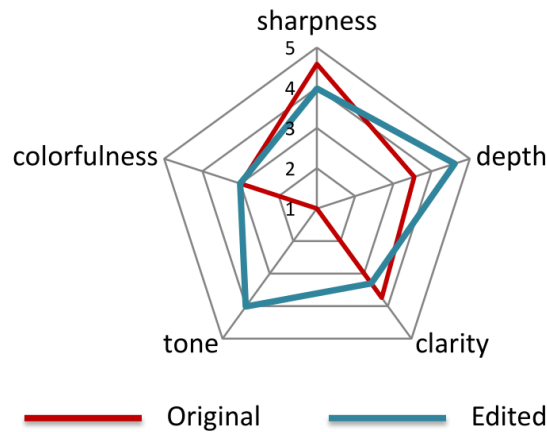


Figure 2.8: An aesthetic signature example with the 5 proposed visual attributes [2].

2.4 Existing Dataset Overview

As an intrinsically subjective task, the collection of reliable data labels is a significant challenge, which greatly impacts the development of effective models. There are two main trends in collecting the aesthetic labels.

The first one is by inferring aesthetic information indirectly from human raters. Authors in [98] collected over 1.7 million photos from Flickr and computed the popularity aesthetic score from each image's number of views and the showing duration. Photo Critique Captioning Dataset (PCCD)[11] collected pair-wise image-comment data from professional photographers from a professional photo critique website. In AROD[89], the authors used the probability of "faves" in "views" as a Flickr image's measurement of the photos' aesthetic appeal. However, in these cases, aesthetics is difficult to distinguish from other subjective values, such as the level of interest, humour, or popularity. These indirect methods are still doubted comparing to the direct aesthetic labels. They are summarised in Table 2.2.

The second one is by eliciting aesthetic quality from subjects directly. However, existing datasets are often limited in terms of reliability, variety or quantity. A summary of the key factors

Table 2.2: Datasets Inferring Aesthetic Information Indirectly

Dataset	Composing Score Information
Sucheckij[98]	views and showing duration
PCCD[11]	critique comment
AROD[89]	faves and views

of datasets who gathered independent image aesthetic information directly are shown in Table 2.3.

The Photo.net dataset and DPChallenge dataset[23, 24, 45] are the earliest datasets for image aesthetic quality assessment. The Photo.net dataset has around 20 000 images with at least 10 ratings for each image, ranging from the least beautiful 1 to the most beautiful 7 in integer, but they only used 3581 images in classifications. Since *photo.net* is a community website for photographers to share their work, many photos are rated skewed to the good levels. DPChallenge dataset collected over 16 000 images that are more diversely rated for the overall quality score.

CUHK [53] and CUHKPQ [70] consist of 12 000 images and 17 613 images respectively, which are labelled in binary as high or low quality. The data of CUHK are also gathered mainly from the DPChallenge website (www.dpchallenge.com), where digital photography contests are held. They preserved the images that have big consensus of the scores. CUHKPQ added some student photos to expand the dataset. The way CUHKPQ collected labels was further improved in [77]. Overall, they fit a limited part of nowadays scenarios.

Later, a series of datasets relate to aesthetic quality are proposed, including Image CLEF[75], Hidden Beauty [87], Kodak Aesthetics dataset [42], YFCC100M [103] and so on. They mainly aimed on the quantity or the applications. Kodak Aesthetics dataset consists more than 1 500 consumer photos, and each one was rated by 4 people on scale 1-100. Hidden Beauty explored using CrowdFlower as crowdsourcing platform to collect labels. They defined the level of aesthetic judgement for 5 ACR scales and had two dummy images. Each image has at least 5 votes. They designed test questions to filter unreliable users, and they chose the test images by three editors. The editors manually annotated aesthetic scores to rank them, and selected images by the inter-rater agreement. However, subjects may disagree with them as beauty score is very subjective. The aesthetic judgement levels are described in words, including professional words and editing technique. This description tried to get rid of the influence of the theme or content, and predefined that low quality is related to "unacceptable" level, and high quality is with "exceptional". These settings may have stopped others using their dataset.

The Aesthetic Visual Analysis (AVA) dataset[77] has been the most popular benchmark

dataset. The voting were collected “in the wild” from DPChallenge website, where photography amateur competitions are held online. It has a large amount of data; it has around 255 000 images from thematic challenges, where each of them are rated by 78-549 different users, in average 210 amateur votes per image; it has 14 style labels based on light, colour and composition, and some of the images are annotated with one or two of the 66 semantic category labels. The general aesthetic scores given by the subjects are ranged from 1 to 10, where 1 is for the least beautiful and 10 is for the most beautiful. However, data collection lacks a precise and properly defined methodology. In particular, as each challenge has a pre-defined theme, when crawling data from online photographic challenges, the aesthetic scores have different interpretations across challenges, and within the competition, voters may have followed different evaluation criteria.

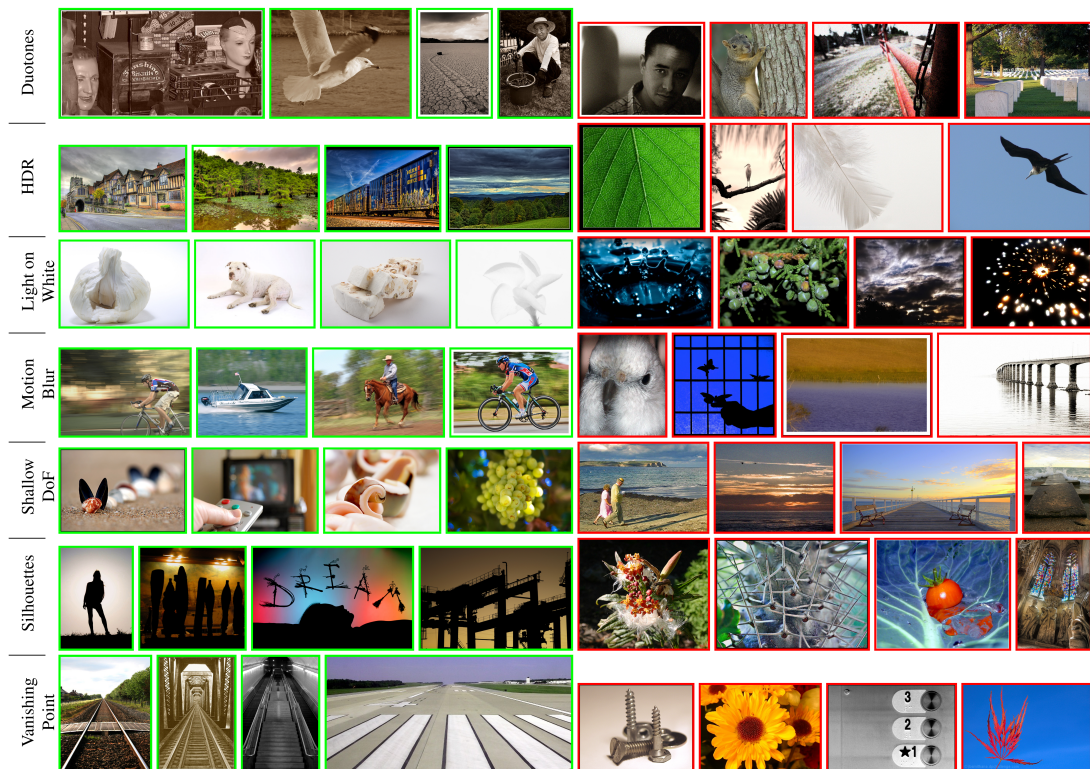


Figure 2.9: Example images in different style categories of AVA dataset [77]. They showed their classification of images in good (green) and bad (red)

Subsequently, the dataset has been expanded by collecting data not only from DPChallenge but also *photo.net* in the Image Aesthetic Dataset (IAD)[68], to meet the needs of deep learning models, but AVA remains the most popular dataset in aesthetic studies.

Other works like [114, 46] crawled data based on AVA to augment information from the users and photographers on the same website. As each subject voted for images in the competitions, some of the users left text comments, which have been collected in AVA-Review dataset[114]. More than 50 000 images and 300 000 comments were collected due to this aim. Conversely,

many photographers' demographic information are collected by AVA-PD[46].

However, these information are not for all the voters and we do not know which voter relates to which votes. It is very difficult to separate the contribution of purely aesthetic factors to AVA scores from any other contextual factor, without knowing the behavior of the original voters. Different from other quality assessment and computer vision tasks, such as object recognition and detection, the ground-truth labels given by human observers in image aesthetic quality remain very subjective and may not follow the Gaussian distribution. In particular, the user training, or lack thereof, can affect the reliability of the final score.

Thanks to the variety of the images, some researchers chose AVA's images and get new votes from external source. For example, the Filter Aesthetic Comparison Dataset (FACD)[99] focuses on filtered images based on AVA dataset. It contains 28 160 filtered images as well as part of the rater labels, and the comparing pairs are done by Amazon Mechanical Turk.

There are also datasets not using AVA images. While they fulfilled the aesthetic datasets, the combination or alignment of the image aesthetics from different datasets are not studied well.

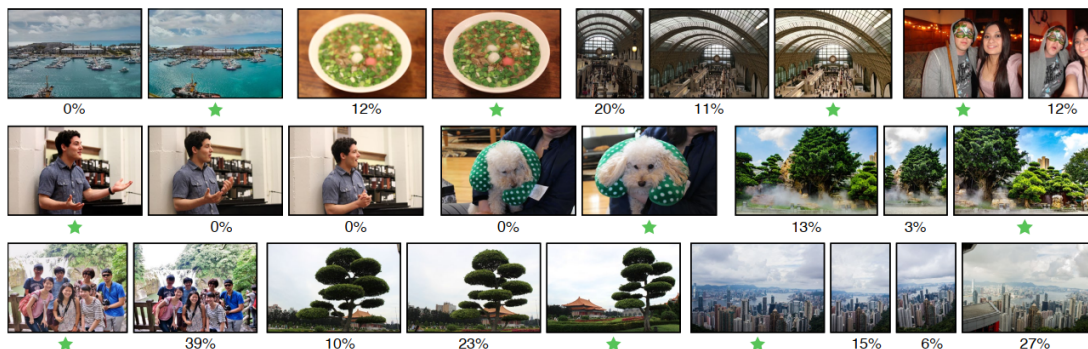


Figure 2.10: Example images of dataset in [10].

Chang et al.[10] collected their own photo album data. They collected near 6000 series of casual photos from college students with a contest, and each series has the same scene. The examples are shown in Figure 2.10. They used Amazon Mechanical Turk to choose the better one in over 90 000 pairs of photos and computed perceptual labels off-line. Each image got a relative rating rank in a series. Similarly, Kuzovkin et al.[58, 59] focused on letting subjects select the best, more representative, or most important photos from real-life scenario photo albums to know their preferences in the contexts, and they mainly fulfilled the data from YFCC100M and other datasets. The ratings are gathered in the context of photo albums, which is different from other independent image aesthetic assessment datasets.

Ren et al. collected Flickr-AES and REAL-CUR [84], where images are downloaded from Flickr, and each image is relabelled by 5 Amazon Mechanical Turk workers. The REAL-CUR is

nominated by letting author rate for their own image album. However, the number of labels for each image is not very convincing on predicting, and their test set and train set do not share users, which is possible to cause a great difference in the judgement criteria. Considering this work mainly focuses on individual aesthetic preference, the dataset still worth mentioning.

The Aesthetic and Attributes DataBase (AADB)[55] tried to record user id and their votes in both general and attributes for 10 000 real photos, and the subjects were asked to choose the level of various attributes directly. The attributes are interesting content, object emphasis, good lighting, color harmony, vivid color, shallow depth of field, motion blur, rule of thirds, balancing element, repetition, and symmetry. They also added style attributes comparing to AVA, and claimed a better balance between consumer photos and professional photos. They applied consistency analysis by comparing Spearman's rank correlation between workers, since the same five workers annotated the same ten images. However, the study was only conducted with 5 or 6 observers per image, recruited using Amazon Mechanical Turk. They voted for the same 10 images. No personal background information of the observers are included, which makes it difficult to make further studies such as impact of user demographics on aesthetics. Training procedures are not provided, so we do not know if the workers understand the attributes and the task, while some attributes like "shallow depth of field", "good lighting" and "color harmony" can cause confusing of non-professional people.

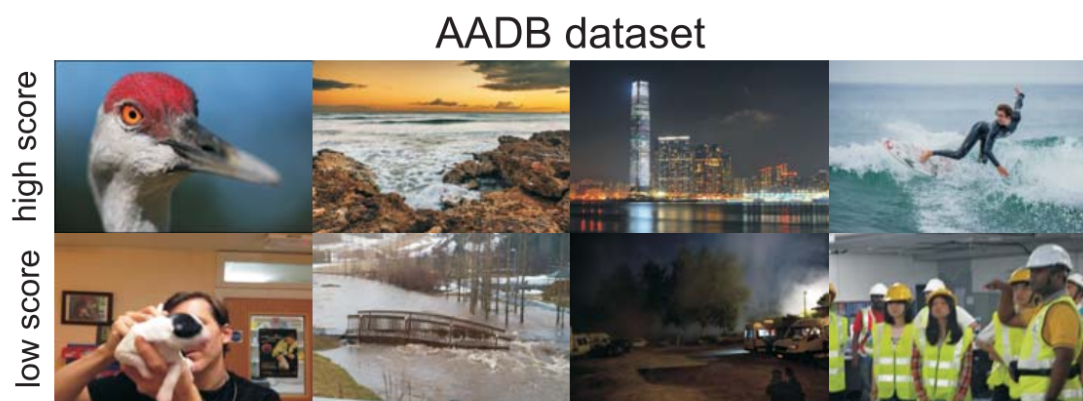


Figure 2.11: Examples of AADB dataset [55].

Images with Aesthetics and Emotions (IAE)[119] is a dataset extension from the authors' previous emotion prediction dataset work, where around 20 000 images are voted by 10 volunteers with 4 aesthetic levels. A combined dataset in [9] from different category's professional photographs is used to analyse aesthetic quality prediction model's robustness, consensus and discriminative ability.

Specific field aesthetic datasets also appeared as the need grows. Portrait photography image dataset[30] is mentioned in their developing of taking a better photo. The authors in [35]

proposed a dataset consists of famous landmark architecture's photographs and 3D models for recommendation a good viewpoint. Food photography images are manually labeled with binary aesthetically judgement and are collected in Gourmet Photography Dataset (GPD)[91].

From the mentioned datasets, we can see that even though there are many different types of dataset, the data are mostly collected in unconstrained conditions, exhibiting noise and making them unreliable.

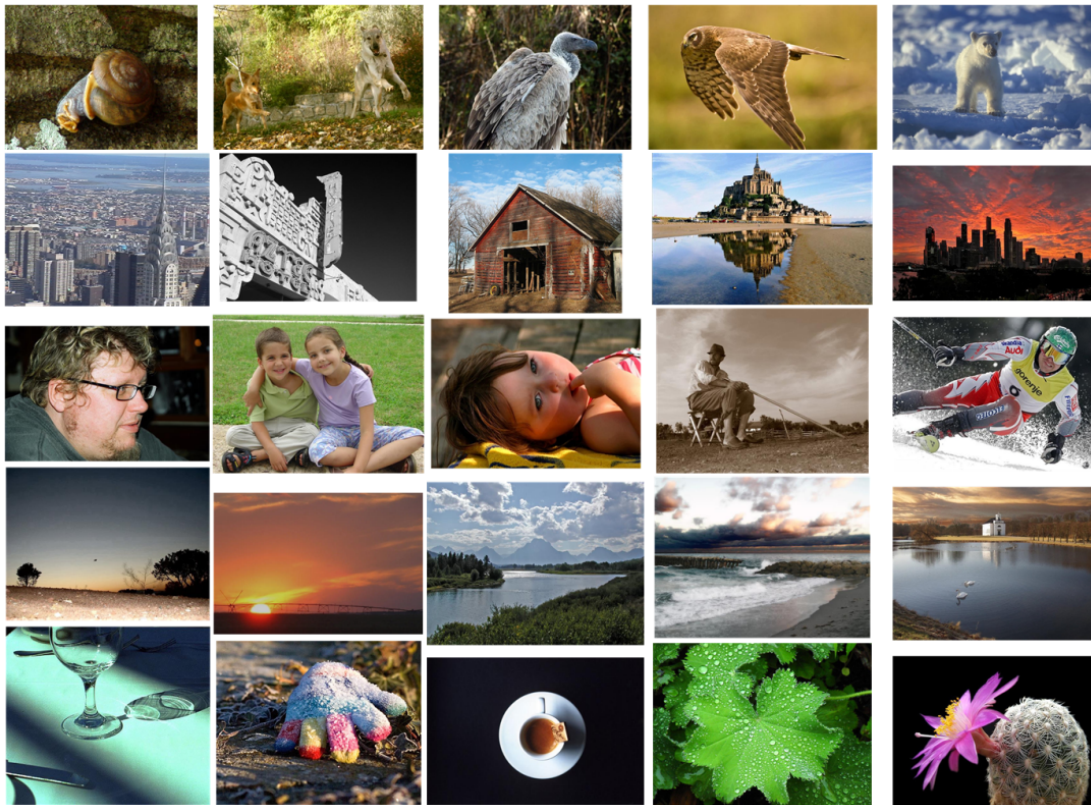


Figure 2.12: Example images from the Waterloo IAA dataset [65]. From the top to the bottom row represent Animal, Architecture/City Scenes, Human, Natural Scene, Still Object images respectively.

Contrarily, authors of Waterloo image aesthetics assessment database[65] collected aesthetic scores in a laboratory environment and controlled conditions. The images in it are claimed to be more uniformly distributed in the aesthetics spectrum than the crawled databases with the same website source. More precisely, they collected users' age (22-33) and gender (18 males and 15 females) information. There is a training but it is only limited to explaining the use of the interface. They trained the users with 20 images to be familiar with the procedure. They used 80 edited images to verify that having a frame of an image is not significantly affecting the aesthetic value, while claiming that all of the 9 tested aesthetics feature are not significantly correlated with the aesthetic scores. They manually labelled their images to five image content types: Animals, Architectures/City Scenes, Humans, Natural Scenes, and Still Object. Though

the detail of this labelling was not given, they gave example images as Figure 2.12. When they selected images, they tried to make them uniformly distributed across types and mean values. They used duplicated images to do the consistency check, since all of the voters voted across the whole dataset. The subjects are with normal corrected vision. They rate in integer for each image on the scale of [0,100], but they were not asked to vote for attributes. Their test lasts 90 minutes with two pauses. It is worth noting that, overall, they only have 33 subjects (26 after removing unreliable users) and 1000 images from *photo.net*. It is obviously not easy to acquire significantly more data under these conditions.

Based on the above overview, existing datasets have disadvantage like noisy-labelled data due to different definitions of the task or with limited training, lack of information for further research, limited scale, reproducing difficult method and so on. We need a new dataset with a reasonable variety and quantity of annotated images where the labels are collected in a disciplined way in order to better explain and study visual aesthetics.

Table 2.3: Overview of datasets that infers aesthetic quality directly without album context. \checkmark means "yes", \times means "not designed" or "not provided". "D" means the aesthetic distributions are provided or can be computed from the public dataset files, "S" is for the aesthetic mean score, "B" is for binary labels. "Website" is for "photo-sharing website". "Lab" is for "lab environment".

Dataset	number of images (\approx)	rating scale	votes/image	subject information	train	attribute label	experiment environment
Photo.net [23, 24]	20k	D, 1-7	≥ 10	\times	\times	semantic label	website
DPChallenge	16k	D, 1-10	\times	\times	\times		website
Kodak [42]	1.5k	1-100	4				
AVA [77]	255k	D, 1-10	78-549	\times	\times	part of challenge, semantic and style labels	website
CUHKPQ [102]	17k	B	10			scene categories	lab
Hidden Beauty [87]	10k	5-ACR	≥ 5	general	\checkmark		crowd-sourcing
IAD [68]	1.5m	D, 1-10		\times	\times	same to AVA	website
Aydin [2]	100	D	21		\checkmark	sharpness, depth, clarity, tone and colorfulness	lab
AADB [55]	10k	D, 1-5	5-6	ID	\times	content interestingness, colour, lighting, focus and composition in 3-level rating	crowd-sourcing
Waterloo IAA [65]	1k	S, 1-100	26	general	limited	content type	lab
IAE	20k	B, 4-ACR	10		\times	emotion	
FACD [99]	1k references, 28k filtered	4 ACR		general	limited	category, filters	crowd-sourcing

Chapter 3

Modeling and Predicting Subjectivity in Aesthetics

3.1 Introduction

Most of existing aesthetic quality prediction methods try to improve the accuracy or prediction error by considering the task as classification or regression, where ground-truth labels are classes or scalars and the algorithms are data-driven. However, while they use single value to describe the aesthetic value of an image, they neglect the intrinsic *subjectivity* of aesthetic quality assessment by human and the ordinal nature of the aesthetic vote to human. We define subjectivity as *the degree of consensus* of the subjects. It can be affected by subject's characteristics in assessment, like personal background, emotion, interestingness, etc. Most of existing datasets checked if the majority of the vote distributions follow Gaussian distribution [77], while they neglected that subjectivity of human votes can cause different shapes of distribution.

Recent research predicts aesthetic distributions. They assume that from predicted distributions, besides the mean score, more information can be computed. However, they did not explicitly define the aesthetic subjectivity [43, 44, 76, 101]. For example, authors in [44] noticed skewness and kurtosis of the vote distributions change for different images, but skewness and kurtosis describe the shape of distribution, and are not adapted for the subjectivity. Jin et al. [43] consider standard deviation as the aesthetic estimation difficulty, which can be computed after predicting the distribution.

Instead, we consider to define explicitly the aesthetics subjectivity of each distribution by a scalar value, for improving the description of rater's consensus and for further improvement on

aesthetic quality assessment algorithms.

Therefore, we proposed several measures of subjectivity based on the distribution of the scores, including simple statistical descriptors such as standard deviation, as well as new proposed features inspired by information theory. We want to evaluate the prediction accuracy of the proposed subjectivity measures using state-of-the-art aesthetic score distribution prediction, and compare these results with directly learning the subjectivity scores.

In this chapter, we first introduce four subjectivity measurements for describing the disagreement of each image's votes in section 3.2. Then, we proposed two schemes: predicting subjectivity directly and from the predicted distributions in section 3.3. After that, we did experiment based on AVA dataset and analysed the result about subjectivity measures in section 3.4. In the end, we tried to use the subjectivity in helping the computer in assessment task to see if the group disagreement can help to reduce the difficulty of learning an image in section 3.5.

3.2 Proposed Measures of Subjectivity

In this section, to describe subjectivity from a distribution of scores directly, we considered two groups of subjectivity measurements. First, we considered statistically motivated descriptors for the consensus. Second, we proposed two measures inspired by information theory.

We consider a dataset of N images $\{I_n\}$, $n = 1 \dots N$, where each image has been voted by M_n human raters on a discrete scale with k levels, $s = \{s_1, \dots, s_k\}$. We model the M_n aesthetic scores x_n for each image I_n as a realization of a categorical random variable with distribution $p_n(x_n)$, which we approximate with the normalized sample histogram $\mathbf{p}_n(x_n)$. Given $\mathbf{p}_n(x_n)$, we define μ_n and m_n as the mean and median of x_n , respectively.

In order to describe the level of consensus of human raters about the aesthetic quality of a given image, we propose using the following measures:

- **Standard Deviation (SD)** of the score distribution, which describes the dispersion of the scores around the average score, that is:

$$SD_n = \sum_{i=1}^k p_n(i) \cdot (x_n(i) - \mu_n)^2. \quad (3.1)$$

A higher value of SD indicates a lower consensus around the average score, and thus higher subjectivity.

- **Mean Absolute Deviation around the median (MAD)**, defined as the sample average

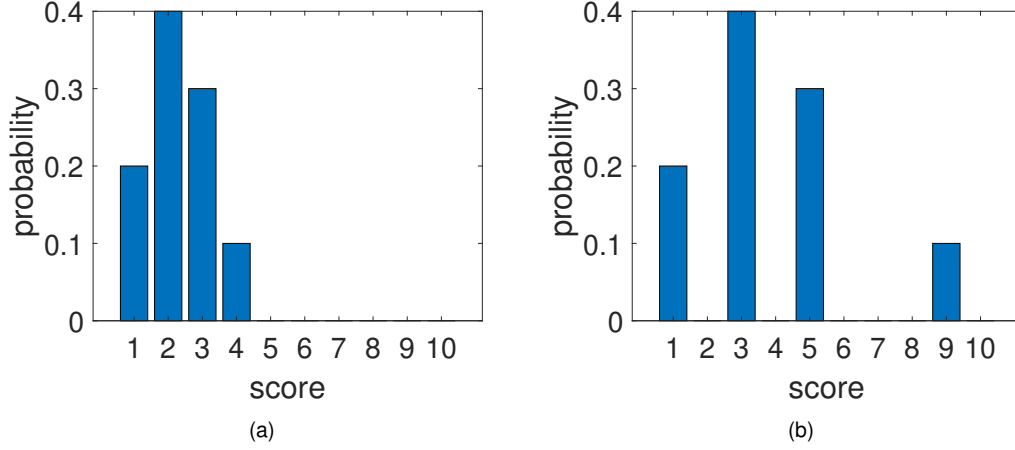


Figure 3.1: The two score distributions have the same entropy, but the one on the right has a higher degree of subjectivity.

deviation of the scores around the median score, that is:

$$MAD_n = \frac{1}{M_n} \sum_{i=1}^{M_n} |x_n(i) - m_n|. \quad (3.2)$$

As for SD, higher values of MAD imply higher subjectivity.

- **Distance to Uniform Distribution (DUD).**

The entropy of a distribution characterizes the degree of uncertainty of the associated random variable, and could be in principle used to quantify subjectivity. However, entropy does not take into account the ordinal nature of aesthetic scores, as illustrated in Figure 3.1. Instead of measuring entropy, we consider the distance of the score distribution $\mathbf{p}_n(x_n)$ from the distribution having the maximum entropy over s , which is the uniform distribution. We quantify this distance using the 2-Wasserstein metric¹ $d_W(\mathbf{p}_n, \mathbf{u}_s)$, that is:

$$DUD_n = d_W(\mathbf{p}_n, \mathbf{u}_s) = \left[\sum_{i=1}^k (\mathbf{P}_n(i) - \mathbf{U}_s(i))^2 \right]^{1/2}, \quad (3.3)$$

where \mathbf{u}_s is the discrete uniform distribution defined over the categories s , and \mathbf{P}_n and \mathbf{U}_s are the cumulative distribution functions of \mathbf{p}_n and \mathbf{u}_s , respectively.

A lower value of DUD implies that the score distribution is more similar to the uniform distribution, and thus the degree of subjectivity is higher.

- **Distance from the Maximum Entropy Distribution (MED).** Since the uniform distribution has always a mean value equal to the midpoint of the score scale, the DUD measure tends

¹Note that the 2-Wasserstein metric is sometimes confused with the Earth Mover Distance, e.g., in [101]. However, for the sake of precision, the Earth Mover Distance corresponds to the 1-Wasserstein metric.

to penalize more skewed distributions having mean values close to the extremes of the quality scale. To overcome this bias, we compare the score distribution with the maximum entropy distribution over the quality scale having the *same mean*. More specifically, we look for a discrete distribution \mathbf{q}_s which solves the following optimization problem:

$$\begin{aligned} & \underset{\mathbf{q}}{\text{maximize}} && H(\mathbf{q}) \\ & \text{subject to} && \mu[\mathbf{q}] = \mu_n, \end{aligned}$$

where H denotes discrete entropy and $\mu[\mathbf{q}]$ is the mean of \mathbf{q} . It can be shown [20] that the solution of this problem is

$$\mathbf{q}_s(s_i) = \frac{e^{\lambda s_i}}{\sum_{i=1}^k e^{\lambda s_i}}, \quad (3.4)$$

where λ is numerically found so that $\sum_i s_i \mathbf{q}_s(s_i) = \mu_n$. Then MED for image n is defined as:

$$MED_n = d_W(\mathbf{p}_n, \mathbf{q}_s) = \left[\sum_{i=1}^k (\mathbf{P}_n(i) - \mathbf{Q}_s(i))^2 \right]^{1/2}, \quad (3.5)$$

where \mathbf{Q}_s is the cumulative distribution of \mathbf{q}_s . As for the DUD measure, the lower MED is, the higher is the subjectivity of an image.

Mean	SD	MAD	MED	DUD
5.174	2.104	1.697	0.278	3.570

(a)

Mean	SD	MAD	MED	DUD
5.180	0.902	0.657	0.650	4.439

(b)

Figure 3.2: Measures that compactly describe subjectivity based on the two score distributions in Figure 1.2 respectively.

The table in Figure 3.2 shows an example of these measures computed for the two images Figure 1.2. We can observe that all of them capture correctly the degree of consensus of the score distributions. In the following, we will study how accurately each of these measures can be predicted, either directly or by means of predicted score distributions.

3.3 Subjectivity Prediction Framework

In order to predict the subjectivity measures proposed above, we consider two options: i) we predict the score distribution *indirectly* using an existing score prediction method; or ii) we compute subjectivity measures on ground-truth scores, and learn to predict them *directly*.

Indirect subjectivity prediction

The underlying motivation of predicting score distributions lies in the possibility to derive aesthetic subjectivity [43, 44, 76]. Therefore, we first consider state-of-the-art aesthetic distribution predictors to estimate the subjectivity measures introduced above, as illustrated in Figure 3.3(a). The advantage of this approach is that, once the distribution is estimated, one can compute any subjectivity measure from it. However, we will show experimentally that this approach is generally sub-optimal compared to directly estimating a subjectivity score.

Direct subjectivity prediction

A limitation of the indirect subjectivity prediction is that the estimated distribution scores are generally noisy, as even the best method to predict the histogram of aesthetic scores has limited performance [44]. In principle, assuming a deep neural network predicting aesthetic distribution approximates the maximum likelihood estimator [33], a subjectivity estimator based on the predicted distributions is asymptotically efficient [51]. However, in practice the number of samples used for training the network is finite, and prediction errors on the distribution might lead to worse prediction performance of subjectivity.

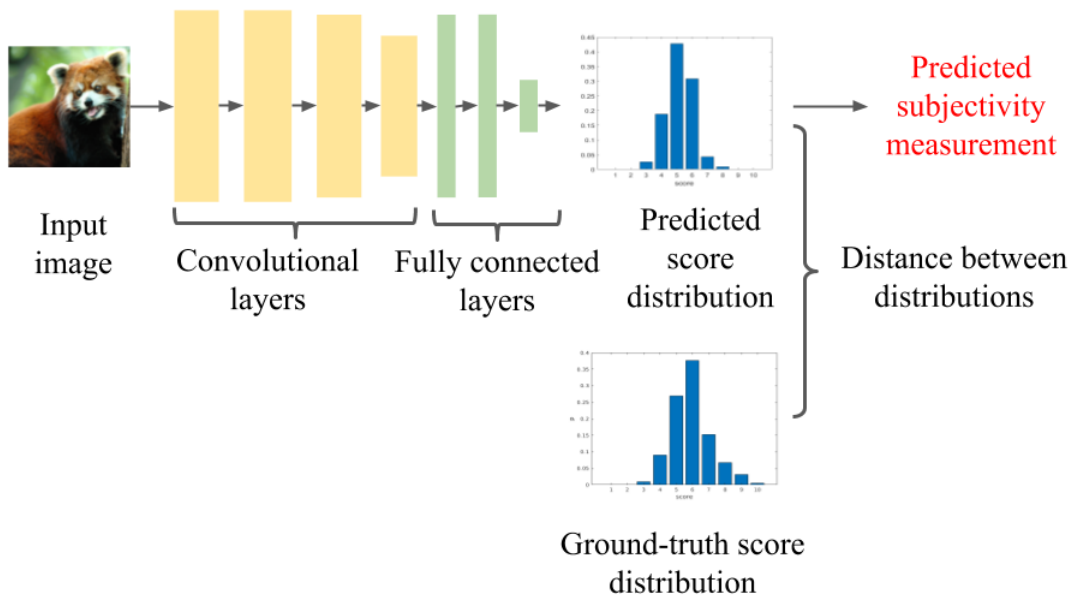
Therefore, we consider the alternative approach consisting in predicting directly the ground-truth subjectivity measures, as shown in Figure 3.3(b). The subjectivity measures are computed on the ground-truth aesthetic distribution. We use afterwards a deep convolutional neural network to predict these subjectivity measures.

3.4 Experimental Results

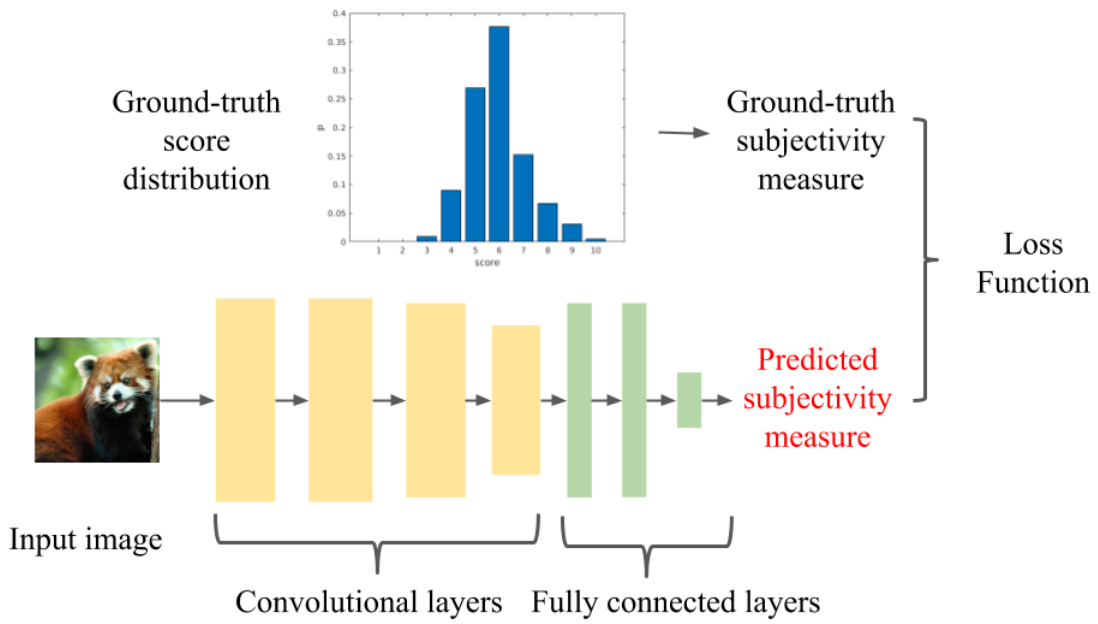
In this section we analyze the prediction performance of the aesthetic subjectivity measures introduced in Section 3.2.

3.4.1 Experimental Setup

We choose Resnet-34 as network structure to predict subjectivity. According to our experiments and the test in [5], Resnet-34 provides similar accuracy result as VGG-16, but uses less memory. In addition, we also use Resnet-101 to study the influence of a deeper network structure in the direct aesthetic subjectivity prediction. The last (fully connected) layer of Resnet-34 is replaced by 3 fully connected layers: two 512×512 fully connected layers plus a 512×1 layer to give a 1-dimensional output. For Resnet-101, the two additional fully connected layers have size 2048×2048 . The drop out rate is 0.5 for every fully connected layer.



(a) Indirect subjectivity prediction



(b) Direct subjectivity prediction

Figure 3.3: Subjectivity Prediction Framework. In the indirect prediction framework, an aesthetic score distribution is estimated first, and subjectivity measures are computed over it. We compare this approach with directly predicting subjectivity computed on ground-truth distributions (b).

We use Pytorch models that were pre-trained on ImageNet [25], and we fine-tune them using training images from the AVA dataset [77]. We use the standard test set of AVA, which consists of 19,930 images. We randomly pick 23,553 pictures for validation, corresponding to approximately 10% of the training set size. All of the input images are resized to 224×224 pixels. Even though previous methods often augment data with horizontal flip, we decide not to do any kind of data augmentation, as differently from classification or recognition tasks, the ground-truth in aesthetics is obtained by human raters and might be influenced by flipping. We employ Adam optimizer [54] and a batch size of 64. The learning rate is decreased by 10 times when the loss does not change over two consecutive epochs. We fix the initial learning rate to 10^{-5} and the maximum number of iterations to 40,000. We employ the L1 norm as loss function for the direct prediction of the subjectivity measures.

For the indirect subjectivity prediction, we consider the following three methods for predicting aesthetic score distributions: the work of Bin Jin et al. [43] (chi-square distance loss); NIMA [101] (Earth Mover's Distance loss); and the RSCJS method of Jin et al. [44] (cumulative Jensen-Shannon divergence loss). Bin Jin et al. provide their trained model (using VGG-16), but use a different (smaller) test set than the standard AVA one. Since their test set is not provided, and for the sake of a fair comparison with other methods, we run their model on the standard AVA test set instead. For NIMA and RSCJS, the original code is not available, and we reimplemented them following the original papers. For NIMA and RSCJS, we use Resnet-34, modified as discussed above.

3.4.2 Performance Indicators

We evaluate the prediction of the subjectivity measures using 4 performance indicators:

- *Pearson's Linear Correlation Coefficient* (PLCC), which measures the linearity of the relationship between the predicted and the ground-truth subjectivity score. Higher values indicate better prediction performance.
- *Spearman's Rank-Order Correlation Coefficient* (SROCC), which indicates the degree of monotonicity of the prediction. Higher values indicate better prediction performance.
- *Mean absolute error* (MAE), which indicates the degree of accuracy. Prediction is more accurate when MAE is small.
- *Mean relative absolute error* (MRAE), which is MAE normalized by ground-truth values. A smaller MRAE indicates higher accuracy.

3.4.3 Experiment Results

Table 3.1: Pearson’s Linear Correlation Coefficient (PLCC)

Methods	<i>SD</i>	<i>MAD</i>	<i>MED</i>	<i>DUD</i>
Bin Jin’s[43]	0.145	0.159	0.178	0.096
NIMA [101]	0.169	0.187	0.211	0.255
RSCJS [44]	0.187	0.199	0.227	0.281
Direct (Resnet-34)	0.274	0.276	0.323	0.351
Direct (Resnet-101)	0.307	0.304	0.333	0.360

Table 3.2: Spearman’s Rank-Order Correlation Coefficient (SROCC)

Methods	<i>SD</i>	<i>MAD</i>	<i>MED</i>	<i>DUD</i>
Bin Jin’s [43]	0.142	0.156	0.162	0.085
NIMA [101]	0.230	0.240	0.250	0.297
RSCJS [44]	0.169	0.152	0.228	0.283
Direct (Resnet-34)	0.267	0.268	0.311	0.351
Direct (Resnet-101)	0.295	0.295	0.316	0.355

Table 3.3: Mean Absolute Error (MAE)

Methods	<i>SD</i>	<i>MAD</i>	<i>MED</i>	<i>DUD</i>
Bin Jin’s [43]	0.294	0.255	0.106	0.226
NIMA [101]	0.171	0.156	0.059	0.122
RSCJS [44]	0.169	0.152	0.059	0.117
Direct (Resnet-34)	0.148	0.133	0.054	0.120
Direct (Resnet-101)	0.146	0.132	0.053	0.101

Tables 3.1-3.4 show direct and indirect subjectivity prediction performance. We observe that direct subjectivity prediction always outperforms indirect prediction through distribution scores, for all the proposed subjectivity measures. In particular, for the same network complexity (Resnet-34), predicting directly the subjectivity is clearly better than predicting the score distribution first and computing subjectivity based on it. A possible explanation can be obtained by looking at the results of distribution prediction, as shown in Figure 3.4, which compares the average predicted aesthetic score distribution vs. the average ground-truth one. For the three distribution prediction methods considered here, we notice that, on average, predicted distributions are different from the original and may even be shifted. Notice that all of the proposed subjectivity measures are affected by errors in the prediction of the histogram.

Although direct prediction improves all the considered performance indicators, we observe that overall the prediction performance is still not satisfactory, e.g., the SROCC is just slightly above 0.35. We might wonder whether this is due to a limited capacity of the Resnet-34 model

Table 3.4: Mean Relative Absolute Error (MRAE)

Methods	<i>SD</i>	<i>MAD</i>	<i>MED</i>	<i>DUD</i>
Bin Jin’s [43]	0.226	0.270	0.210	0.053
NIMA [101]	0.129	0.162	0.128	0.029
RSCJS [44]	0.127	0.158	0.127	0.028
Direct (Resnet-34)	0.107	0.130	0.126	0.025
Direct (Resnet-101)	0.104	0.130	0.120	0.024

we employed. Therefore, in order to study how subjectivity prediction performance improves with a more complex network, we tested the direct prediction scheme using Resnet-101, which is much deeper than Resnet-34. As expected, the results generally improve over the simpler Resnet-34. However, this improvement is in most cases only marginal, showing that aesthetic subjectivity prediction is intrinsically a hard problem – at least a harder one than predicting the average aesthetic score, where SROCC between predicted and ground-truth values is higher than 0.6 [101].

Comparing the different subjectivity measures, those inspired by information theory (DUD and MED) are in general those with higher prediction performance. Among the statistical motivated descriptors, the SD is generally predicted more accurately than MAD. We can assume that, for the same neural network model complexity, a ground-truth variable which has a higher dependence on the input is easier to predict, or, in other terms, target variables which tend to be more “noisy” will be more difficult to learn. Thus, we can argue that the subjectivity measures based on information theory are somewhat more robust than statistical deviation measures. A possible rationale behind this could be that both DUD and MED are based on distances between histograms, which take into account the whole score distribution. On the other hand, SD completely captures data variability when the underlying score distribution is Gaussian, which is the case for only 62% of AVA images [44]. MAD is supposed to be more robust to skewed distributions, but it might be affected by the sample median computation, which on a 10-dimensional distribution as for aesthetic scores can only take values over a small set, i.e., $\{1, 1.5, 2, \dots, 10\}$.

Notice that the DUD measure achieves the best correlation among the four subjectivity measures, despite the fact that it penalizes more those images with distributions having mean score far from the midpoint of the quality scale. These are also the images that are less frequent in the AVA dataset. Therefore, DUD might implicitly act as a weighting scheme during learning, similar to [43], where weight is the inverse of frequency of MOS and more frequent MOS is penalized in training.

3.5 Improvement of Mean Aesthetic Score Prediction

3.5.1 Method

In this section we study how to use subjectivity information to improve the estimation of aesthetic scores

Images whose score distribution present higher consensus among human voters are those for which mean aesthetic score prediction is generally more meaningful. In other terms, a larger aesthetic score prediction error can be more tolerable when the ground-truth score distribution has less consensus, since humans in general would not agree on the aesthetic value of the image. On the other hands, when the consensus is high, the same prediction error might lead to more misleading conclusions about the image aesthetic value. Furthermore, we can expect that more attention should be given during the training to those images whose aesthetic mean score is more reliable, i.e., the aesthetic subjectivity of those images is lower.

Therefore, we propose to use subjectivity to weigh the error of mean score prediction. When subjectivity is high, the mean score is not well representing the "true" aesthetic score of the picture. Different from [44], which observed that kurtosis values of distributions have different distances to 3 (where 3 is the kurtosis value of a normal distribution) when the shapes are tailed differently, and used kurtosis as a measure of reliability, we use our proposed subjectivity measures as weights. Specifically, we penalize more the images with higher consensus among voters, in order to learn more accurately their mean aesthetic score.

In training, the loss of each learning epoch decides the learning direction. If a mean score has a big subjectivity, then this mean score is less reliable to describe this image. Thus, we would expect the images that are more difficult to be agreed by the group have a less importance in the loss function. Therefore, we use the subjectivity to compose the weight, as in function 3.6:

$$WMSE = \frac{1}{\sum_{i=1}^N 1/(s_i + \theta)} \sum_{i=1}^N 1/(s_i + \theta) \cdot (y_i - \hat{y}_i)^2 \quad (3.6)$$

where s_i is the SD or MAD of the image's aesthetic general score distribution. θ is a parameter to avoid denominator to be 0. For SD and MAD, the bigger s_i is, the less people agree about the score, and the smaller importance it will take in the loss function.

For MED and DUD, since the smaller they are, the bigger the disagreement is, the equation 3.6 should be changed to 3.7:

$$WMSE = \frac{1}{\sum_{i=1}^N s_i + \theta} \sum_{i=1}^N s_i \cdot (y_i - \hat{y}_i)^2 \quad (3.7)$$

where s_i is the MED or DUD value. In practice, no subjectivity measure in AVA dataset has a

Table 3.5: Predicting Aesthetic Score with Weights

Methodscriterion	MSE	SROCC
Resnet-34 without weight	0.339	0.624
Resnet-34 with SD	0.337	0.628
Resnet-34 with MAD	0.347	0.613
Resnet-34 with DUD	0.335	0.630
Resnet-34 with MED	0.322	0.649

value equal to 0, thus we set $\theta = 0$ for both loss functions.

3.5.2 Experiment and Result

In the experiment, we used the previous standard test set provided by AVA.

The training network structure is as in Figure 3.5(a), which is a Resnet-34 network with fully connected layers replaced by two 512×512 fully connected layers plus a 512×1 layer to give a 1-dimensional output. The pre-processing procedures are the same. The inputs are the resized image, the image's aesthetic mean opinion score, and the subjectivity value. We used the pre-trained Resnet-34 network provided by Pytorch. The learning rate is set to be 10^{-4} , and the learning rate is decreased by 10 times when the loss does not change over two consecutive epochs. The batch size is 64, and the maximum number of iterations is 20 epochs. We employ the Mean Square Error loss function. For the testing part, the structure is as shown in Figure 3.5(b), and only the image is put into the network.

The results are shown in Table 3.5, where the first column shows the structures and subjectivity measures we used, the second column is the Mean Square Error (MSE), and the last column is Spearman's Rank-Order Correlation Coefficient (SROCC).

We can see that adding subjectivity as a weight of score can improve the score prediction's correlation coefficient. The measures inspired by information theory works better than the ones motivated by statistics. SD gives correlation coefficient as 0.628 and MSE reduces from 0.339 to 0.337, comparing to directly training the aesthetic score. MAD gets a negative effect, decreasing the SROCC from 0.624 to 0.613, and the MSE rises from 0.339 to 0.347. DUD slightly works better than SD, reaching to 0.630 in correlation coefficient and 0.335 in MSE. The most improving measure is MED. Its MSE reduces to 0.3222 and SROCC rises to 0.649. Since the correlation coefficient values are close, we use methods in [124] to test if adding subjectivity makes a difference. The four confidence intervals of the difference of two correlation coefficients(adding a subjectivity measure and Resnet-34 without weight) do not contain zero, so it can be deduced that the adding subjectivity is significantly different not adding it.

To this end, MED is the most effective subjectivity measurements in AVA in helping to predict

the aesthetic mean opinion score. It is reasonable because comparing to statistical methods, DUD and MED considered the ordinal nature of distributions, so they can distinguish the degree of consensus in more cases, e.g. Figure 3.1. Comparing to DUD, MED considered mean score of the distribution in measuring the distance to max entropy distribution, so it penalize different mean scores more equally in loss function.

Looking at the relationship between the subjectivity measures and the mean opinion score computed from AVA dataset, shown in Figure 3.6, four measures show different behaviours. For lower SD values, which means people voted for similar scores, the mean opinion score sits around the medium score. Medium consensus images have various of MOS. When people have the least consensus, MOS is closer to medium scores in the range. MAD has similar behaviour, but with smaller MAD values (<1), the score ranges wider than the one of SD. DUD has the smallest value around medium score as well, and when it increases to the extreme values, there are two poles of MOS. Contrary to DUD, MED has two poles of MOS in lowest values, and has highest values around the medium score. This shows that extreme aesthetic level image's MOS has a bigger possibility of being more reliable and meaningful than the medium aesthetic level image's. Thus only predicting mean score or predicting a distribution without accurate subjectivity is not enough.

The fact that high consensus images have a MOS around medium score could be due to the fact that when people are voting for an image in a set of values, when they feel it is difficult to decide, they tend to choose around the medium score effortlessly.

According to study in [38], a quadratic shape can be observed in many subjective tasks. However, comparing to the quadratic shapes illustrated in the work, the square of standard deviation in Figure 3.7 follows a different shape in AVA dataset, but still keeps some quadratic shape. However, since images in AVA dataset have different number of votes, we can not verify if the number of votes is affecting the shape. This shows the complex subjective nature of image aesthetic quality assessment.

Over all, if we use subjectivity value directly as a weight of aesthetic score, we can have improvements, but we are not able to improve the prediction significantly. Even though MED gives the best performance in AVA dataset, which subjectivity measure is the best in all of the natural images is still doubted. We need more well-labelled data in both quantity and quality to study into the difference among measures in practice.

3.6 Conclusion

In this chapter, we have analysed the problem of defining and predicting aesthetic score subjectivity, intended as the degree of consensus human raters express about the aesthetic value of a picture. To this end, we have considered several measures of subjectivity, and two possible subjectivity prediction frameworks.

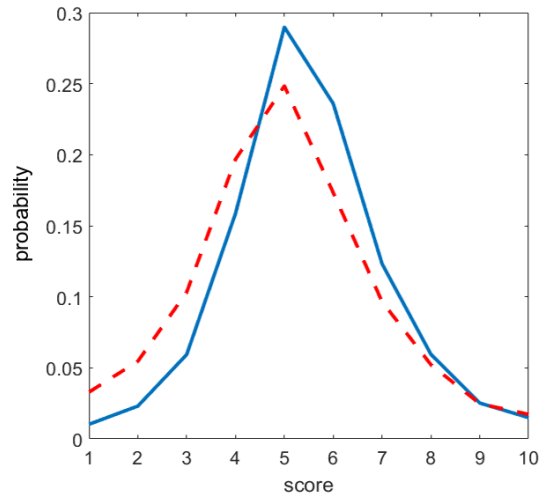
Among the analysed descriptors of subjectivity, we have found that our newly proposed measures inspired by information theoretical principles are, in general, easier to learn, indicating that they might be more discriminative and robust compared to simpler statistical deviation measures.

In helping predicting the aesthetic score, adding subjectivity measures brings limited gains. The measures inspired by information theoretical principles showed relatively bigger gains on predicting the aesthetic values in AVA dataset.

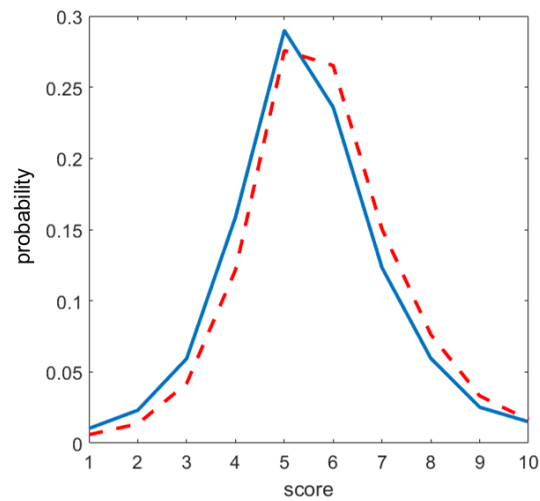
We have shown that predicting subjectivity from predicted score distributions is, in general, sub-optimal compared to directly predicting it from ground-truth subjectivity scores. This indicates that, in practice, aesthetic score distribution predictors are not sufficiently accurate to enable assessing correctly the aesthetic subjectivity.

Despite our approach achieves state-of-the-art subjectivity prediction performance, we recognize that predicting subjectivity is a much harder task than predicting the average aesthetic score. Not only histogram-prediction-based methods can achieve correlations of 0.6 or higher for that task, but also directly predicting mean aesthetic score can reach a good correlation.

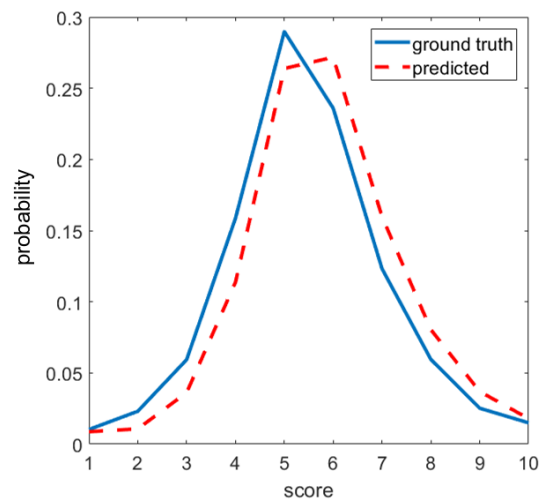
We believe that this is partially due to the complexity of the task in itself, and more importantly, to the noisy nature of current aesthetic datasets. This is evident for the benchmark AVA dataset, where aesthetic scores are influenced by many factors that go beyond the pure aesthetic value of a picture, and the collected method is not controlled enough. Limited information and noisy labels make it difficult to explain what leads to higher consensus with similar mean aesthetic values.



(a) Bin Jin's

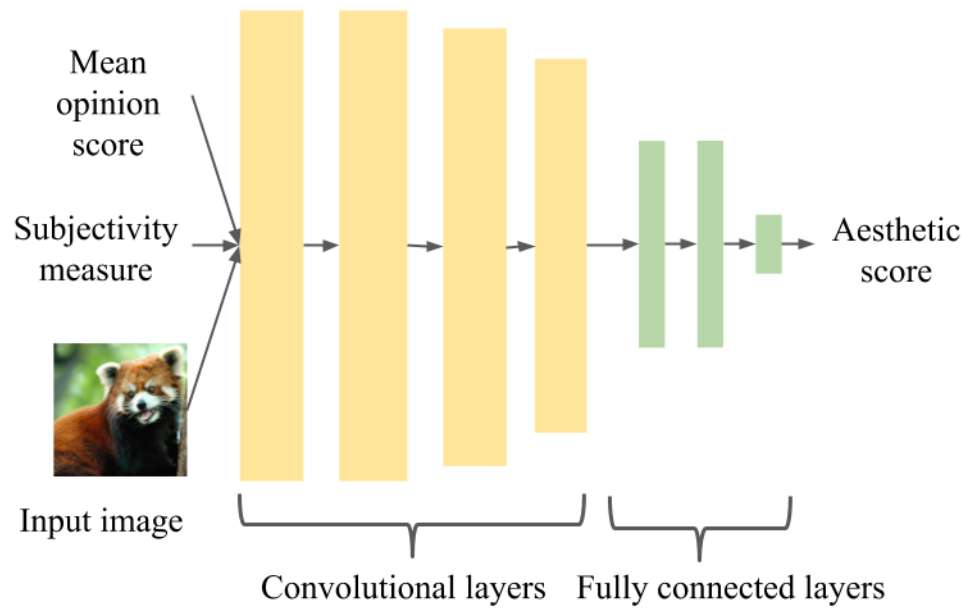


(b) NIMA

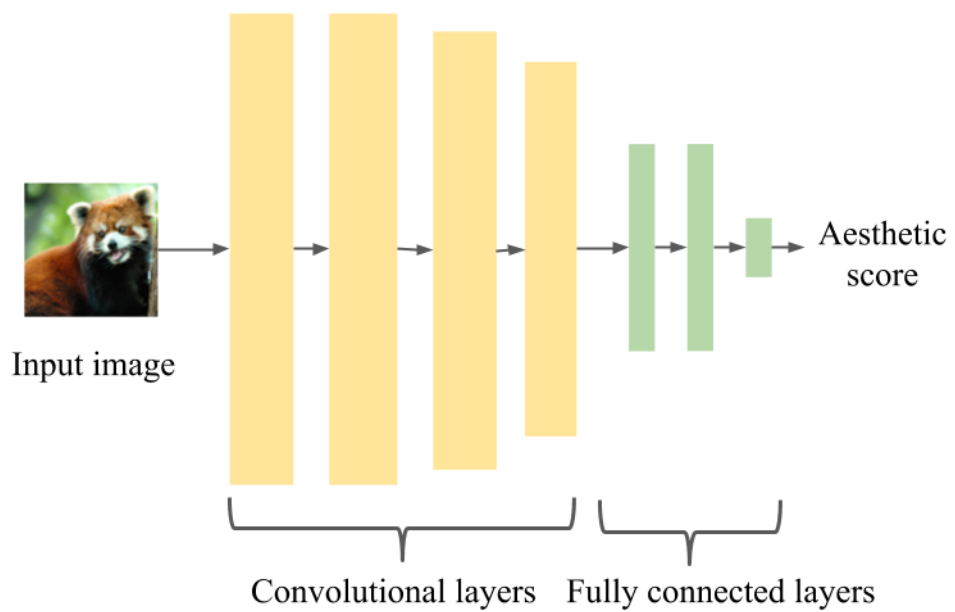


(c) RSCJS

Figure 3.4: Average predicted score distribution vs. ground-truth score distribution over the test set, for the three considered state-of-the-art distribution prediction methods. Notice that for all of them, the average predicted distribution is shifted compared to the original.



(a) Training



(b) Testing

Figure 3.5: Network structures of training and testing

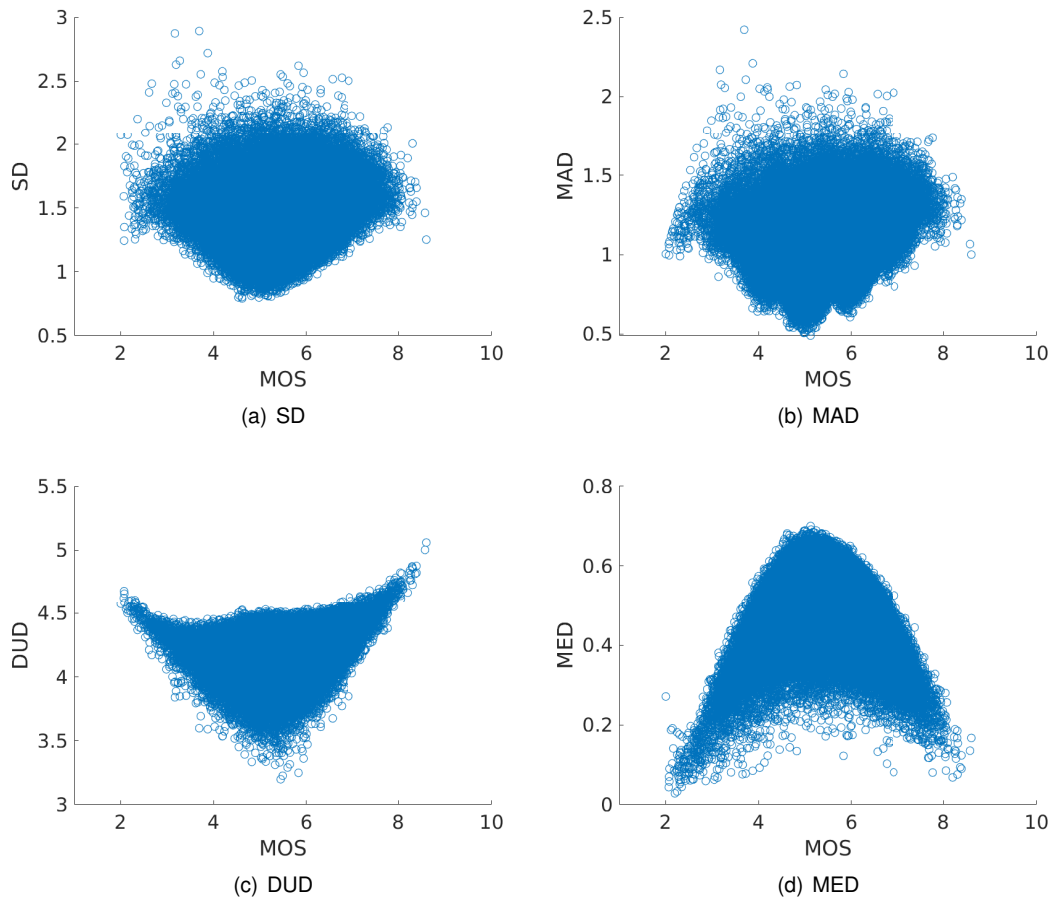


Figure 3.6: AVA subjectivity distribution. Higher SD and MAD indicates a lower consensus, and higher DUD and MED means a higher consensus.

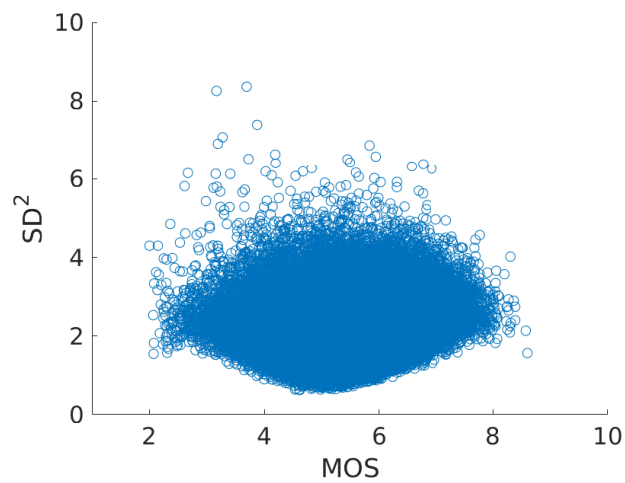


Figure 3.7: Relationship between the square of standard deviation and MOS in AVA dataset.

Chapter 4

EVA: an Explainable Visual Aesthetics dataset

In this chapter, we present in detail the methodology for collecting data in the proposed Explainable Visual Aesthetics (EVA) dataset. First, we describe the whole process. After that, we discuss about the selection of test images from AVA dataset. After collecting data, we removed unreliable data, and gave a brief summarise of the data.

4.1 Work Flow, Platform and Experiment Settings

The general work flow for each user is described in Fig. 4.1, where the circles mean "start" and "end", diamonds represent decisions and rectangles represent processes. The users were told that we were using their cookies to create accounts recording their votes. The whole process is anonymous. Observers first need to provide background information including year of birth, region, gender, and whether they are color blind or wearing glasses. Then, they have to indicate, by their self-assessment, their experience in photography, as either beginner (without any specific knowledge about photography); intermediate (a casual photographer without specific training); or advanced (having followed some specific training in photography).

After registering this background information, observers have to undergo a training phase

Table 4.1: Level of photographic experience

Level	Description
Beginner	I do not have any specific knowledge about photography.
Intermediate	I am a casual photographer without a specific training.
Advanced	I have followed some specific training.

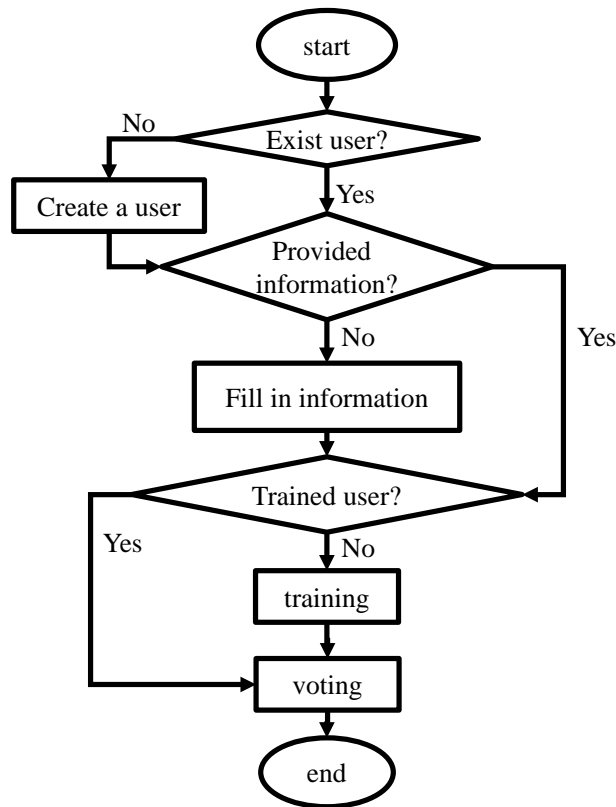


Figure 4.1: General Flow of Survey

in order to better understand the test. More precisely, each survey question is explained and sample images (not present in the test stimuli) are shown for illustration. This step is especially important to ensure that different naive observers receive the same instructions and understand the meaning of each attribute they will have to assess. To verify that they have carefully read the instructions, observers have to click several check boxes intertwined with the text.

The screenshots of the training phase can be seen in Figure 4.2 to Figure 4.6.

After the registration and training phases, observers can start voting, as detailed in the next section. Images are randomly selected for display, and each image is only displayed once for an observer. The images were resized automatically to fit the display width of the device. In order to stabilize judgements, the first two images are dummy stimuli, i.e., their scores are discarded. For the sake of flexibility, subjects are allowed to leave voting at any time and come back later, while being identified with the same cookie account. Since this task is relatively difficult than objective tasks and the tasks that only ask for a general aesthetic score, to encourage the subjects to focus in one session, a counter indicating the number of voted images is displayed, and 30 is considered as a milestone in the counter. Voters are allowed to see their own 10 favourite images, so that they have the willing to continue voting.

Instructions for Voting


Please read carefully the following instructions before starting the test.

General task

The goal of this survey is to assess the aesthetic quality of an image. You will be asked to vote 7 questions for each given image.

1. What is the overall aesthetic quality of this picture?

1. What is the overall aesthetic quality of this picture?



Least beautiful 4 Most beautiful

You need to pull this slider, and you will see the score under it.

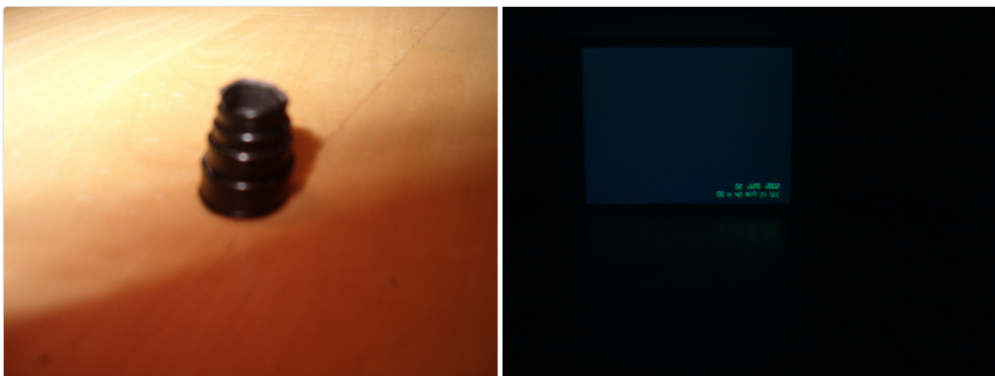
How beautiful does this picture look like? This is not only about the semantic content of the image, but also about its visual characteristics (composition, quality, etc.). More about this will be asked in some of the questions below.

For example, most people tend to find the following images more beautiful:



(a)


and the following images less beautiful:



In the dataset, you will see images that span the whole range of the voting scale (0-10). 0 means the least beautiful; 10 means the most beautiful; 5 is the medium value.

2. How difficult is it for you to judge this image's aesthetic quality?

2. How difficult is it for you to judge this image's aesthetic quality?



Very difficult 5 Very easy

(b)

Figure 4.2: Training's screenshot, part 1.

For example, how long did it take for you to decide the overall aesthetic quality of this picture?



An example of "Very easy"

"From my point of view, this is very easy to be judged as 9/10. I give this score instantaneously."



An example of "Difficult"

"I would give it 4/10, but I'm not sure. It took me a while to judge."

3. How do you like this attribute?

We divided the factors that may affect your judgement into 4 attributes.

A. Light and colour

This question relates to visual perception. It includes several aspects, for example:

(a)

Brightness



High brightness

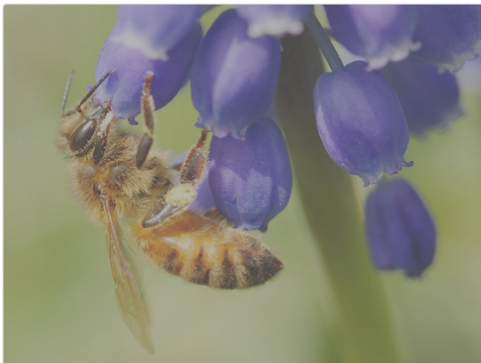


Low brightness

Contrast



High contrast



Low contrast

(b)

Figure 4.3: Training's screenshot, part 2.

Colour Saturation

Saturation of a colour corresponds to how far a colour is from grey.



High saturation



Low saturation

Have you read this part? no yes

B. Composition and Depth of Field

(a)

These images have difference in composition



(b)

Figure 4.4: Training's screenshot, part 3.

These images have difference in depth of field



Have you read this part? no yes

C. Quality

Image quality is the outcome of many factors. Here are common ones:

Blur



(a)

Artifacts



(b)

Figure 4.5: Training's screenshot, part 4.

Noise



Original image

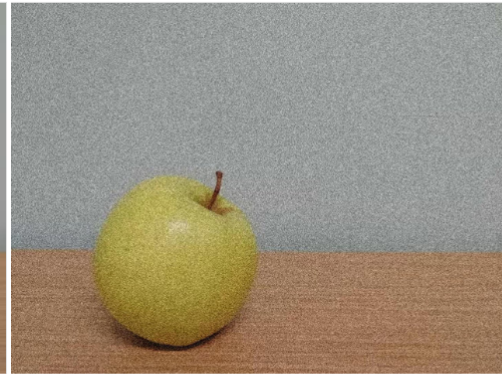


Image with noise

Have you read this part? no Yes

D. Image content

This question relates to the how much you like the content of the image.



(a)

Figure 4.6: Training's screenshot, part 5.

While voting, user behaviour is recorded. More precisely, we record the time when a subject submits each vote. Then we can compute the time laps between two votes. In particular, this information is useful to identify when a subject is voting very fast. It could also be potentially related to the difficulty of assessing a given image. The device type of the users are also recorded. Due to the privacy concern, the exact size of the device screen is unknown.

The experiments reported in this thesis were carried out on-line from February 2020 to July 2020, participated by volunteers over 16 years old. To get more data as well as diversity, two web hosts have been used, the first one in English is hosted outside of mainland China, and the second one in simplified Chinese is hosted in mainland China. The translation was double-checked by other researchers who speak Mandarin and English fluently and a translator, to not only make sure the translation does not make misunderstandings and changes of the questions that we are working on, but also both professional users and non-professional users can understand the questions.

Due to the size of the collected dataset and the limited available budget, and also we would like to collect users' information, we did not resort to any recruitment platform (such as Amazon Mechanical Turk and Zooniverse) to enroll anonymous participants without personal information in this study. Instead, the study was advertised through our social networks, targeting mainly acquaintances, colleagues, scientists and students in vision-related topics. The volunteers were invited to vote for at least one session (30 images) or more. The users that voted for a large number of images were rewarded with some small gifts if they apply for them. Most of the votes come from France and China, reflecting our geographical location and personal relationships.

Besides, users can check the training instructions in a static page, in case the beginners want to recheck the attributes.

For the design of the back end, the two web hosts exchange information everyday, as shown in Figure 4.7. As illustrated in the blue part, each host has a table of the copied votes from another host, as the black arrows, and it also has a table of its own users' voting information. The two tables in each host consist of a view of all of the votes, as the white arrows show. We summarise all the information from two hosts in a local database in a computer, so that we can do other works like checking users' behaviours and knowing the progress, as shown in the off-line green box.

While voting, the images are divided to several groups with ordered group numbers. Each group has approximately 500 random images from the images we selected, so we have 10 groups of images in database. Our goal is to let each image reaches at least 30 votes. Thus, when a user starts to vote, he will see a random image that does not reach to 30 votes in an unfinished group and that has not been voted before by the same user; if there is no unfinished

group, then he will go to the group in the next order. Dummy images are randomly selected in the images that people have voted for 30 votes, so that users would not see them in their valuable voting. Since there is a delay in synchronization of the two hosts, there are some images that got more than 30 votes because there were many users that voted these images in the same day in both hosts.

Because of this synchronization, even though the users are anonymously allocated numbers as user id, we add "E" before the users using English web host, and add "C" before the users using the Chinese web host. This does not mean that users in the English web host does not read Chinese, nor the users in the Chinese web host does not read English. The difference is mainly the geographic difference while we did the experiment.

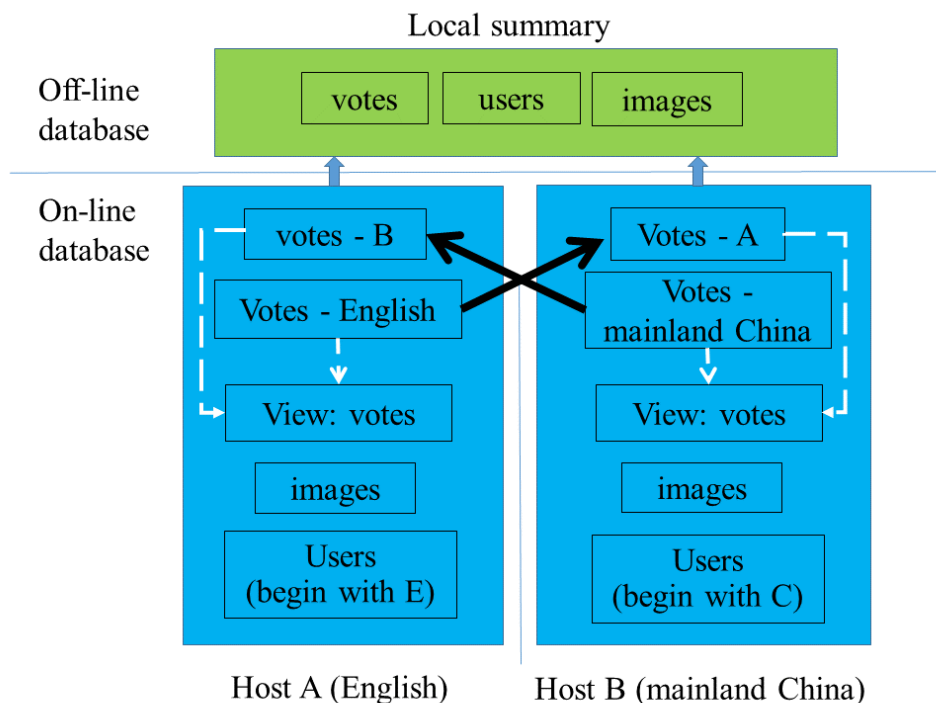


Figure 4.7: Synchronization of two web hosts

4.2 Survey Design

Considering the attributes in previous work [77, 55, 65, 2, 47] and inspired by methods for subjective quality assessment experiments in laboratory conditions [65], we design the survey considering four main attributes and one measure of difficulty to judge image aesthetic quality. More specifically, the survey is composed of several questions and detailed hereafter:

Question 1: "What is the overall aesthetic quality of this picture?" We employ an 11-point

discrete ACR scale. We choose to use discrete ACR scale instead of continuous scales and binary judgement considering its high reliability in image aesthetic appeal study [95], as high reliability means high agreement exists in the ratings that the different users gave to the same image. It leads to a smaller MOS confidence interval. However, instead of the usual categories (excellent, good, etc.), we label the extremes of the scale as "least beautiful" (corresponding to 0) and "most beautiful" (corresponding to 10), and let subjects rate through a slider bar. When a new image is displayed, the slider default position is always set to 5.

Question 2: "How difficult is it for you to judge this image's aesthetic quality?" Subjects have to select an option over a four-level Likert scale: very difficult, difficult, easy, and very easy. Indeed, aesthetic quality is very subjective, and sometimes it is difficult to assign a score to an image. We set the number of options in the Likert scale to be even to avoid the possible tendency of voters to select effortlessly the middle, neutral option. While it is obvious that the consensus on the overall aesthetic score varies significantly across images, predicting the subjectivity is a difficult task [47]. The purpose of this question is to directly ask the subjects about the difficulty to score a given image, with the objective to support further studies on aesthetic subjectivity.

Question 3: "How do you like this attribute?", where we consider four attributes: *light and color*, *composition and depth*, *quality*, and *semantics* of the image. For each attribute, subjects have to vote on a four-level Likert scale with the following options: very bad, bad, good, and very good. We choose these four attributes, as they have been previously studied [50, 107, 55, 2] and they are relatively easy to understand by naive subjects. The attributes have been defined as follows in the user training phase prior to the test:

- Light and color: it relates to visual perception, including brightness, contrast, and color saturation.
- Composition and depth: it relates to the position and spatial relationship between objects in the scene.
- Quality: it can be impacted by different types of distortions, including blur, compression, noise, and other artefacts.
- Semantics: it is related to how much the subject likes the content of the image.

Notice that these attributes span different levels of factors affecting image aesthetics, from perceptual (light/color and visual quality), photographic technique (composition, depth) to higher level features of the scene. We purposely keep the number of attributes to 4, without further detailing them (in particular for composition and semantics), to avoid complex categorization

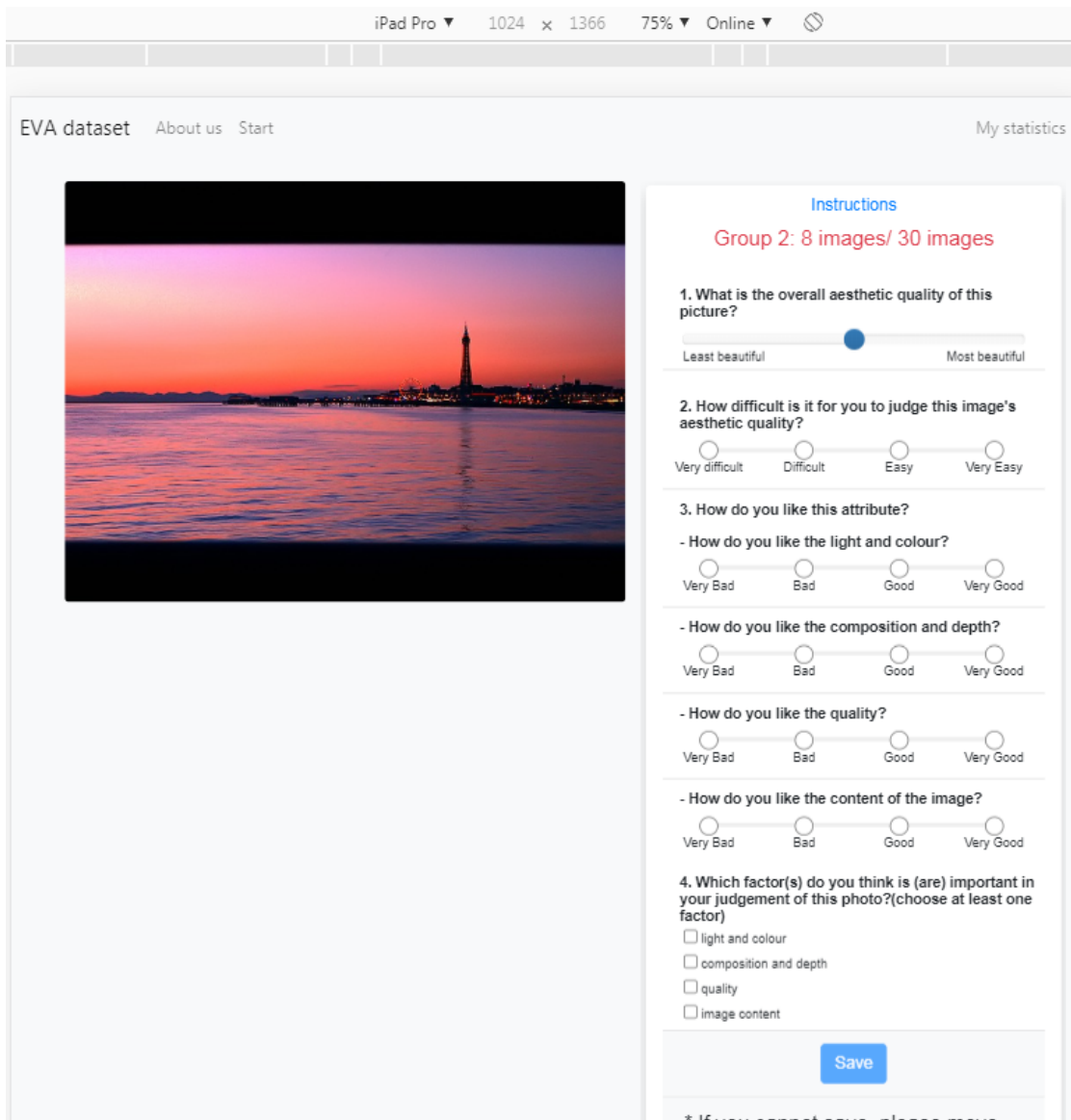


Figure 4.8: Voting page's screenshot on mobile devices. When the user pulls the bar in the first question, there will be a score under the bar. If one of the 4 questions is not answered, the user cannot trigger the "Save" button, and there will be an alert.

which might require more advanced photographic knowledge as well as longer training/test time.

Question 4: "Which factor(s) do you think is (are) important in your judgement of this photo? (choose at least one factor)" where people can choose multiple options among the four attributes mentioned above. The subjects are required to vote for at least one option. We set this question as binary check boxes, in order to avoid making the voting time for an image being too long.

Figure 4.8 shows the voting page's screenshot in tablet and other mobile devices.

4.3 Image Selection

We select the stimuli for our dataset from AVA dataset. AVA is well-known for its very large size and variety. Moreover, many research works have built upon AVA, for example, [46] augments AVA data with photographer information (AVA-PD). We select images which are present also in the AVA-PD dataset, in case people want to study the relationship between photographers and image aesthetics based on our dataset. Our goal is to select more than 5000 images in total, with the procedure highlighted below.

Since the semantic content can influence aesthetics [102], we choose images from different content classes, so that the users are shown different photography categories with a similar probability. Inspired by the 5 content types mentioned in [65], we divide images into 6 categories: animals; architectures and city scenes; human; natural and rural scenes; still life; other, which means none of above. In order to get a rough categorization of the test images, we use Yolo V3 [83] to detect and classify the objects in images. We manually group the object labels given by Yolo V3 model (which has 80 pre-trained classes) into our first five categories. Then, we assume that the category of an image is defined by its main objects. For this purpose, we compute the cumulative surface of the detection frames corresponding to each category. If this surface is over 50% of the entire image or larger than the total surface of the other objects, then this image is classified in the associated category. Otherwise, it is considered in the "other" category. The latter case may therefore correspond to images with several significant objects or no significant object.

Noticing that this is for selecting images from AVA dataset. It is not shown to any of the subjects in rating, and since Yolo V3 can only guarantee the 80 classes' classification accuracy, this is a rough classification procedure. Other more sophisticated classification approaches might be used in further studies.

Existing classifiers are limited in deciding photography content themes. Yolo V3 is chosen to do roughly classification because of its acceptable accuracy and speed. Some images are difficult to classify for photographers as well. We have three people to ensure the validity of the classification. They first looked at image examples in some photography classification tutorials online and how image classification works for computers, then they check images one by one. If they do not agree on an image, it is put into "Other" category. For example, a boy playing with a dog can be put into "Other" because it is difficult to say it is purely human or animal; a model figure's photo, because it can be visually considered as a human and a static object; a strong blurred image can be put into the category because it is difficult to judge by the content. In general, "Other" category contains the images that they cannot classify by photography theme

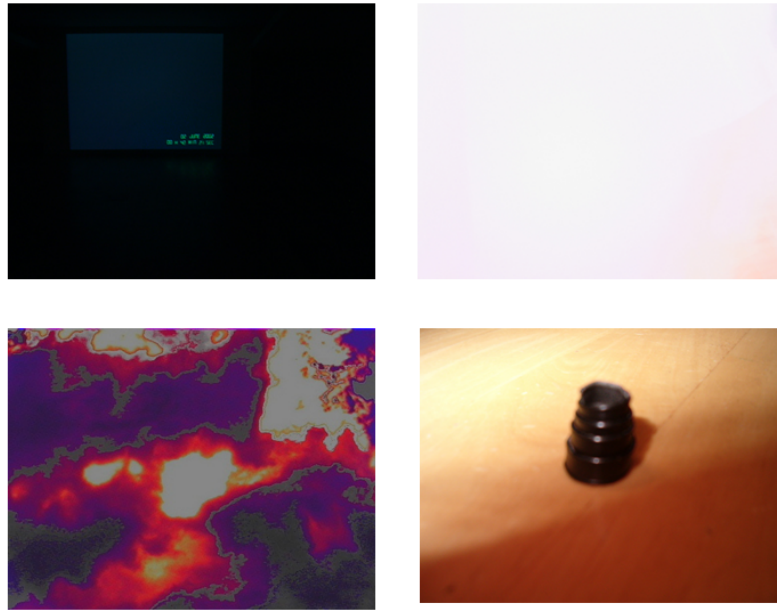


Figure 4.9: Images in AVA dataset with MOS under 4

clearly and confidently and the images that cannot be defined with a single label. Off course, there can be other judgement in classification. For our experiment, it is just for analysing the data in the future, so the users are not shown this information at all.

To take into consideration potentially harmful or uninteresting content, following common academic ethics and administrative measures, images with specific characteristics are also removed. Grounds for removal include sexual content, religion and political sarcasm, drugs, and horror. For the obvious disturbing photos that we are not considering in this thesis, artwork, images comprising a lot of text, and commercial advertisements are also removed to reduce the noise.

A few examples of each category are shown in 4.11. The final photography category is also recorded as an attribute label for each image.

In the AVA dataset, the images with very low quality scores typically have poor technical quality, as a few examples shown in Figure 4.9. Compared to the time when the AVA dataset images have been collected, nowadays the technical quality of photo sensors and imaging system is greatly improved, and even low-end smartphone cameras are capable of capturing pictures with little noise, blur or compression artefacts. Therefore, in EVA we choose to consider only images with reasonable technical quality, as those with very little quality can be easily detected nowadays with existing methods [101]. Moreover, in this way, one could learn a more precise predictor of aesthetics over a smaller range of aesthetic qualities, which might be more useful in an image recommendation or enhancement scenario.

Therefore, we selected images from AVA with associated scores within the range [4,9]. More precisely, we divided the scores in four intervals: [4,5), [5,6), [6,7), [7,9), and selected a similar number of images randomly in each group.

Based on the above procedure, we have selected 5101 images, which are nearly evenly distributed in terms of Mean Opinion Score (MOS), as illustrated in Figure 4.10.

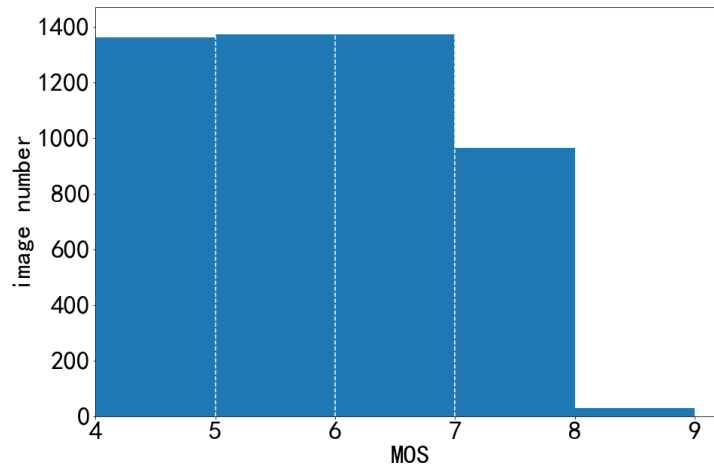


Figure 4.10: AVA score distribution for the images in EVA.

4.4 Data Quality Control and Cleaning Procedure

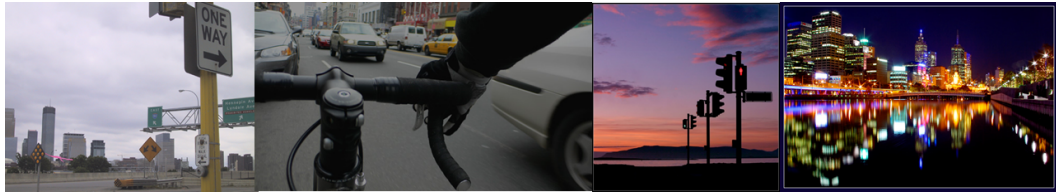
After collecting the data, we get 4734 voting sessions. Of these, 1251 voting sessions contain at least one image after the two dummy stimuli. Only a quarter of visitors finished dummy stimuli probably because first, it requires user information, which some users would not like to provide; second, around half of the subjects did not read the training procedure carefully or patiently so that they did not pass to the next step; third, it is a volunteer work with effort in time and mind comparing to other subjective tasks, so people are not so motivated.

Dummy images are the images subjects voted after the training, with the same routine and user surface, but the result is not included in the dataset, to make the users get used to the task. The users would never vote them again. Notice that the same individual person might have voted in different voting sessions, if the latter are far apart in time, as the cookies expired after a few weeks of inactivity.

In designing the experiment we did not include online quality controls or trap questions. On one side, given the subjective nature of aesthetic scores, it is difficult to detect whether a vote deviating from the average is due to a malicious behavior or simply due to a personal judgment. On the other hand, other kinds of controls such as content questions [95] might be used; however, these can be easily circumvented as users learn to anticipate them when voting



(a) animals



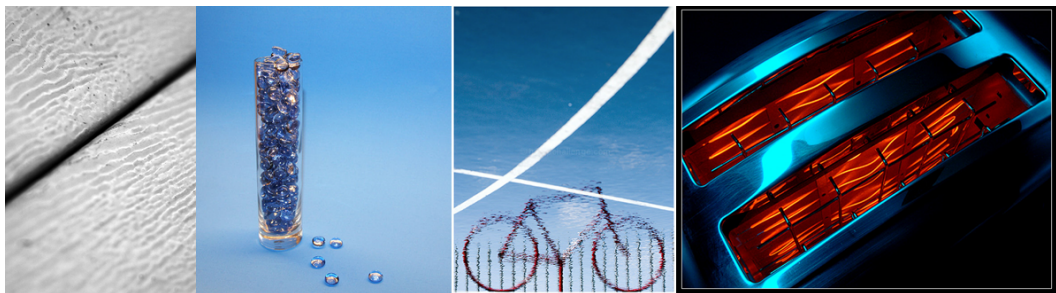
(b) architectures and city scenes



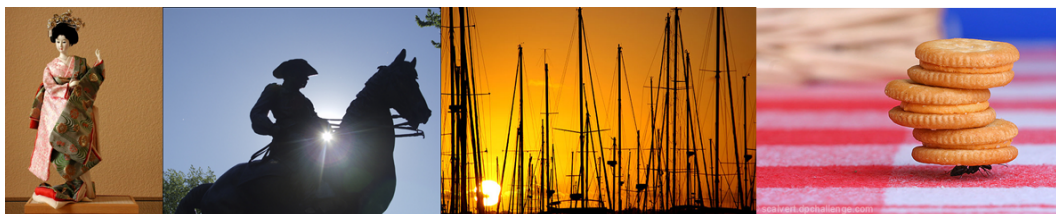
(c) human



(d) natural and rural scenes



(e) still life



(f) other

Figure 4.11: Image examples in each category for EVA

many images, since we allow the users to leave when they feel tired. There is a possibility that users know about the question pool quickly. We observed that tested users got uninterested in doing our test quickly in two of our pre-tests. In one test, they were asked if there is a dog in a photo after 10 votes, and in another test, they were shown a photography trick after voting for 10 votes. Neither of them had a good result and made the users stop after finishing 20 votes, maybe because it is much easier than our questions, or it disturbed the subject's paces in voting. Thus, we relied on the personal engagement of the voters, most of which volunteered the task. We carefully checked the votes of potentially suspect participants, by monitoring constantly the evolution of the votes and eliminating those we deemed to be unreliable. This was indeed a time-consuming activity during the dataset collection.

When we recruited volunteers, we send emails to groups, and also advertisement on WeChat, universities and colleges. To the strangers, they were told that they could get small gifts if they tell us their id and let us check their data immediately plus a short talk with them. At the beginning, we found some users constantly voted one score under 4 for images and came for applying for small gifts. By asking about the explanation of the votes, they admitted that they were voting randomly to get the gifts. Some users did not understand that we wanted the data to study instead of to finish a quantity, and they voted a lot in a single day. In experiment of [39], they also observed that many workers are voting for a high frequency for any single answer choice in 5 scale ACR. They manually set a threshold of the portion of same choices to filter the workers. As we have 11 scale, we can remove these subjects by using standard deviation as a criteria. In these cases, the user with all his/her votes were removed from the data pool. Then, people were told that we were able to find outliers. In a few student groups, the outliers were pointed out with the anonymous id, which made others intuitively get serious. Besides, we added a warning of exceeding 300 images in a day. If the users voted more than that, they were told the votes would not be recorded but they could continue, and the values are not written in the database. The images they saw after this threshold would not appear again in future voting. These phenomenons disappeared quickly by doing so. We have this manual check from database about the outliers every two days, and synchronized the databases in two web hosts everyday.

In total, 172934 raw votes (before data cleaning) have been collected. Due to the use of a crowd-sourcing data collection approach, the gathered data may include outliers, which need to be identified and removed. We apply statistical a posteriori analyses to filter out the collected votes. However, while some inter-rater agreement indicators [34] such as Cronbach's alpha or Intra-Class Correlation (ICC) have been proposed for aesthetic subjective analysis [95], the high number of votes collected in EVA required that each image is evaluated in general by a different

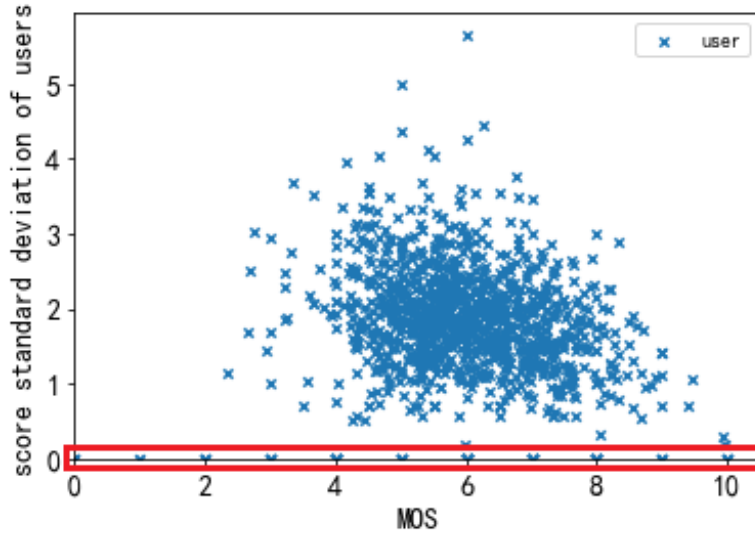
combination of users. As a result, inter-rater variability is difficult to evaluate and interpret. On the other hand, post-filtering approaches such as CrowdMOS [85], which compares the consistency of individual votes with the population, assume Gaussianity of the image score distribution, which is generally not the case for aesthetics. Therefore, we consider alternative approaches to post-filtering the votes, but we release also the raw data of EVA to allow further analyses and data cleaning methodologies to be employed in future research.

In our analysis, we employ the recorded voting time associated to each individual vote as an indicator of possible under-commitment of users to the task: voting times that are too short might imply that a voter assigned scores randomly [95]. Specifically, we obtain the voting time as the time interval of two consecutive votes. In order to collect reliable statistics about the minimum voting time, we identify a group of 13 trusted voters, including users that have previously participated to other lab-based user studies organized by the authors. By inspecting the distribution of voting times for this pool of users, we observed that the minimum voting time is 7 seconds, which we select as a threshold on the minimum voting time to consider a vote valid. About 3% of the votes in the dataset correspond to a voting time smaller than 7 seconds, and are then discarded as possible outliers. Other thresholds of cleansing by voting time can be done according to the needs with the dataset.

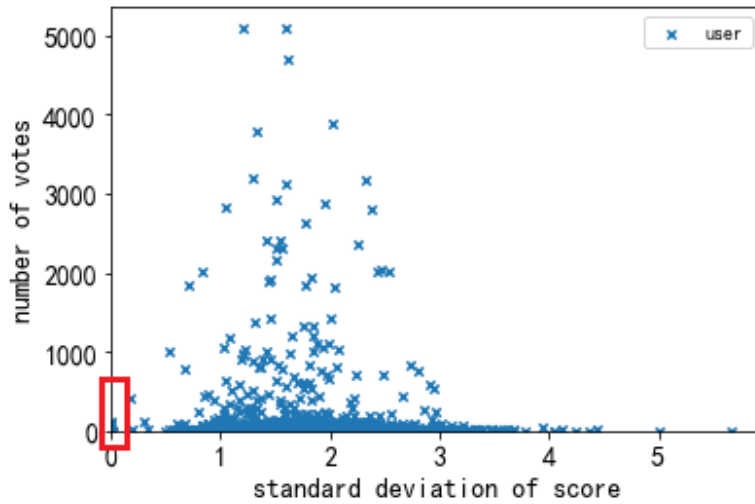
As a second criterion for outlier detection and removal, we consider the standard deviation of votes for a user. More specifically, we observe that voters with very small standard deviation in their judgment of global aesthetic score (e.g., those who gave the same vote to all images they voted) might be unreliable. Figure 4.12(a) shows a few users have a standard deviation close to zero and their MOS is not gathered at the same value. Figure 4.12(b) shows these users have few votes. The sessions with a standard deviation under 0.1 are distributed across the whole range of [0,10], while the standard deviation a bit higher than 0.1 are not evenly distributed. It is difficult to say that the latter ones are voting randomly or their discriminate ability is limited since we do not have ground-truth.

Thus, we empirically set a threshold of 0.1 on the standard deviation of an individual user's votes, in order to decide whether he/she is an outlier. We removed 156 voting sessions with a standard deviation under 0.1 from the dataset, and their vote number distribution is shown in Figure 4.13. Notice that most of these voting sessions actually consists of very few images, so the impact over the whole dataset is rather limited.

To have a robust estimation of aesthetic scores and similar number of votes in each image, we remove from the dataset images having less than 30 votes. After this data cleaning process, 4070 images have been retained, with 30 to 40 valid votes each, as distributed in Figure 4.14.



(a) Each User's MOS and Standard Deviation of Votes



(b) Each User's Number of votes and Standard Deviation of Votes

Figure 4.12: Each user's general aesthetic assessment's mean score, standard deviation and number of votes. The red frames show the suspected outliers.

4.5 Data Summary

In this section, a brief summary of EVA dataset is given. The cleaned dataset includes 4070 images, with a total of 136943 votes from 1094 voting sessions.

Figure 4.15 reports statistics about the participants of the study. Around 30% of users use computer or laptop, and the rest uses mobile devices (including smartphones and tablets). Females and males have a similar population. The majority of the subjects come from mainland China, the second most is coming from France, and the rest come from other places. Most subjects wear glasses; a small amount of people have color blindness. Most people know little about photography, a part of subjects are amateurs, and a small amount of voters have

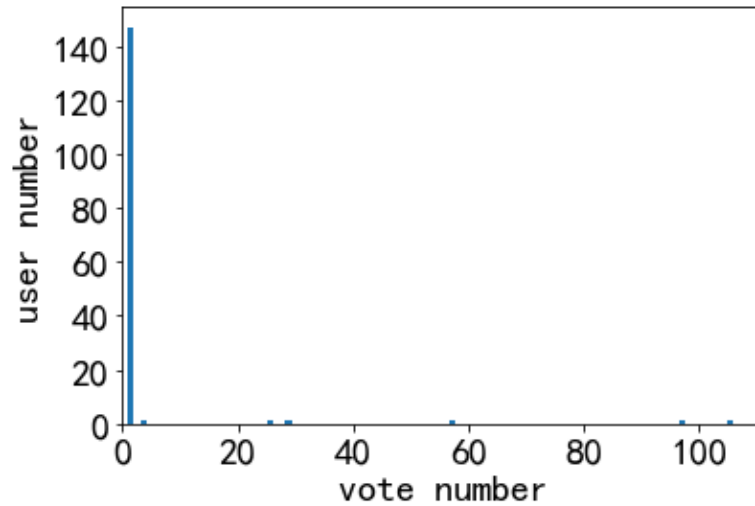


Figure 4.13: Distribution of vote number of users whose voting general aesthetic scores' standard deviation is smaller than 0.1

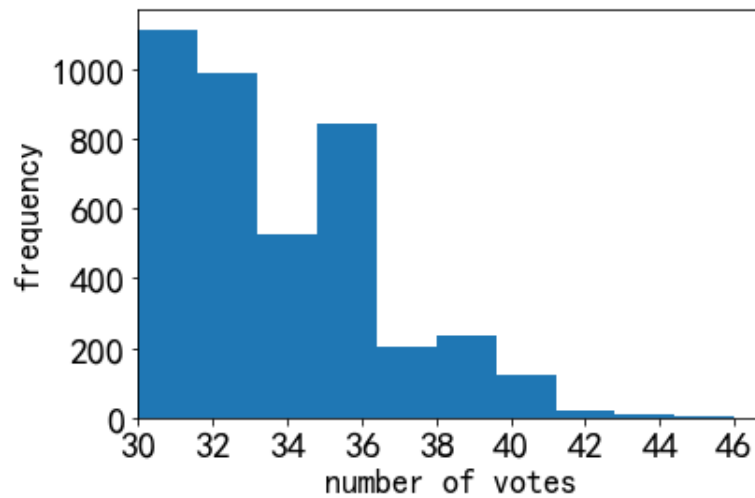


Figure 4.14: Number of votes for each image

professional skills. This likely reflects a realistic distribution of photographic skills across the population and is an intended feature of the EVA dataset, which targets aesthetic perception at large.

The average aesthetic score distribution is illustrated in Figure 4.16 and 4.17. The highest peak for MOS is around 6, rather than the medium score 5, probably because the images have relatively high quality. The peak for standard deviation of scores in each image is between 1.5 and 2.0. A Shapiro-Wilk test [81] performed on these distributions reveals that they are not Gaussian distributed.

The number of each user's votes is demonstrated in Figure 4.18. There are people voted few, either because their voted images are removed because of lack of votes, or the subjects were tired to continue. The median vote number is 27.

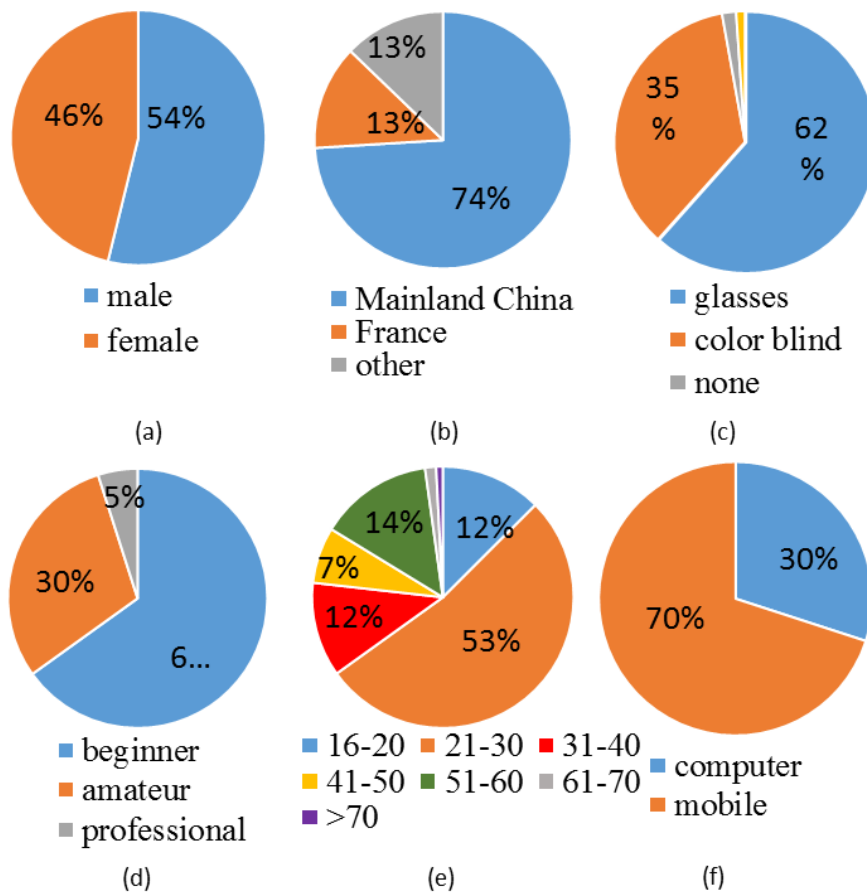


Figure 4.15: Statistics of votes in EVA dataset: (a) Gender (b) Region (c) Visual Status (d) Photography Experience (e) Age (f) Device

For attributes, the distributions are shown in Figure 4.19. We can see that they are skewed, and the peaks are around 3.0, which means "good". This may be due to the way images have been selected from the AVA dataset, i.e., images with very low quality have been discarded.

Figure 4.20 show the result of the last question, about the summary of the importance of factors that the subjects think. We first computed the sum of the choosing times of each factor in each image, then we normalised by each image. Then, we compute the frequency of each attribute getting the most important attribute in an image. In the end, we normalise the frequency. We can see that the composition and depth is considered to be the most important attribute among the four, and color and light is slightly less important. Quality is the least important above all. Figure 4.21 shows the importance to the subjects in different content categories.

Table 4.2 gives a brief comparison to other datasets popular for deep learning methods. Comparing to crowd-sourcing environment AADB and lab experiment environment Waterloo IAA dataset, we combined their advantages and overcame some disadvantages in building

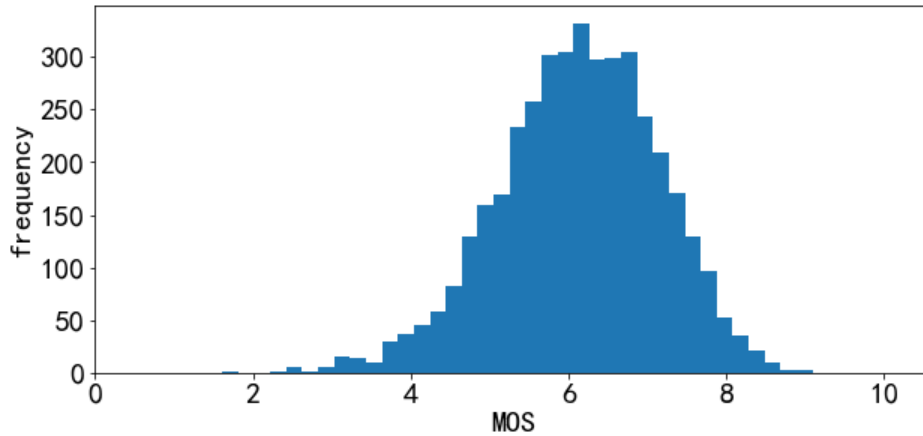


Figure 4.16: Distribution of Mean Opinion Score (MOS)

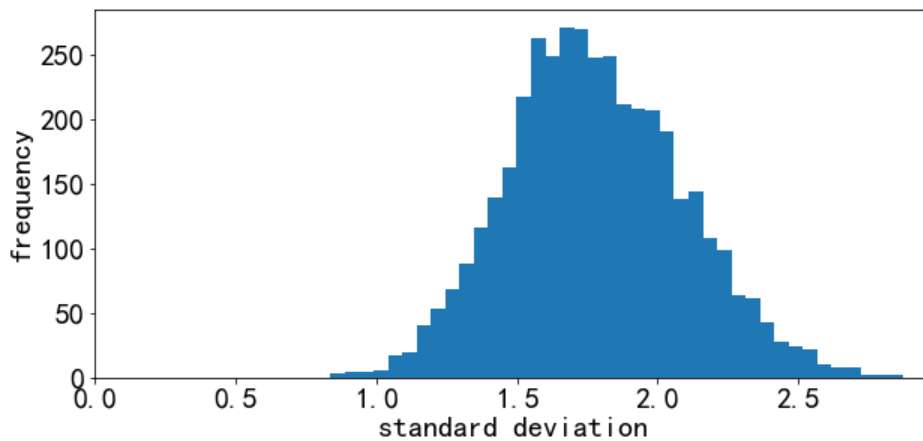


Figure 4.17: Distribution of Standard Deviation (STD) of scores

EVA. We collected aesthetic distributions and various labels in EVA, which can be useful for explaining aesthetics. We considered that AADB only asked each 5-6 people to vote for a batch of 10 images, the rating ranges of aesthetics and attributes are small, so it risks in introducing too much bias; Waterloo IAA dataset only collected around 1000 images, while their training is only limited to explaining the use of the interface, and the experiment environment setting limits the subject amount, variety and the speed of collecting data. Thus, we chose crowd-sourcing method to collect more variety and quantity of data. We selected images considering different aesthetic levels and content. To reduce the noise in judging, we added a training of the concepts for the subjects.

As the benchmark dataset for many deep learning methods, AVA has a huge amount of images, but AVA contains edited images, advertising and many other images that probably share different assessing standards, and collected labels that were given in competition context without identifying users for data cleaning. Thus, AVA may have introduced noise from the beginning. EVA has a smaller quantity of data regarding the time limit, but it collected votes from

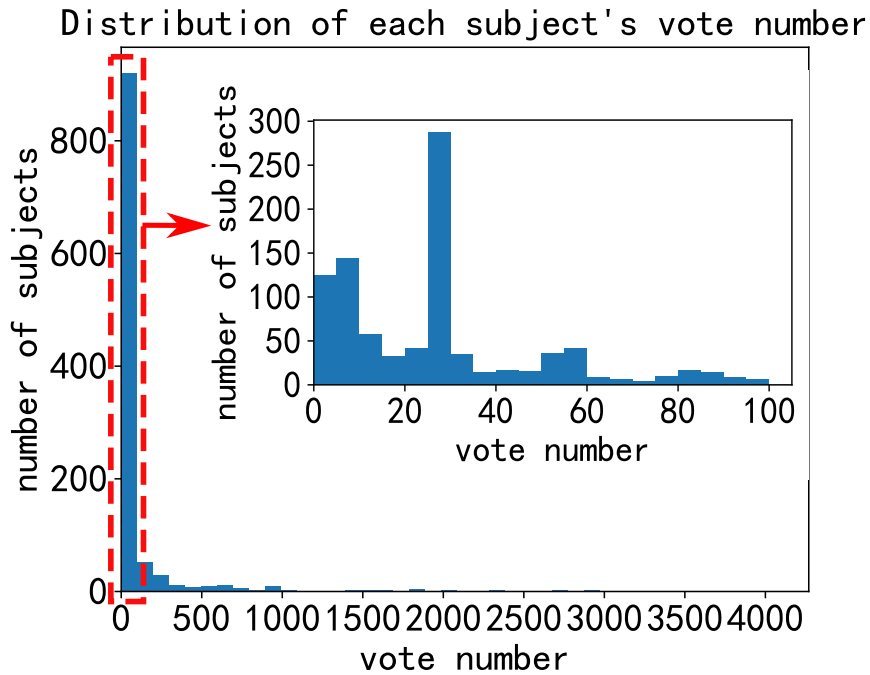


Figure 4.18: Distribution of user's vote number

an independent image assessment experiment, where images were pre-filtered and randomly shown without reference to the subjects. It has training before the assessment, so the users can share a more similar understanding of the questions, and EVA collects subject's voting data for data screening. Therefore, EVA's noise is more likely to come from the subjective task and the subjects.

Since we selected images from AVA dataset, we can compare their general aesthetic score as illustrated in Figure 4.22, 4.23, 4.24. Comparing to AVA dataset's labels, EVA's labels have a bigger range, but the quantity of low aesthetic quality (< 4) is small. The portions of high aesthetic quality (> 7) are similar in two datasets. We can see that some images share very different opinions in two datasets, and the majority share the same tendency.

4.6 Conclusion

In this chapter, we described how we collected the first annotated image dataset for explaining visual aesthetics in a controlled practice. It contains 4070 annotated images with 30 to 40 votes per image, collected using a disciplined approach including subject training and unambiguous definition of aesthetic attributes inspired by traditional quality assessment guidelines. As a result, EVA overcomes the limitations of previously proposed datasets, in particular noisy labels due to misinterpretations of the tasks or limited number of votes per image. At the same time, it offers a number of novel features, including the degree of difficulty in judging the aesthetic

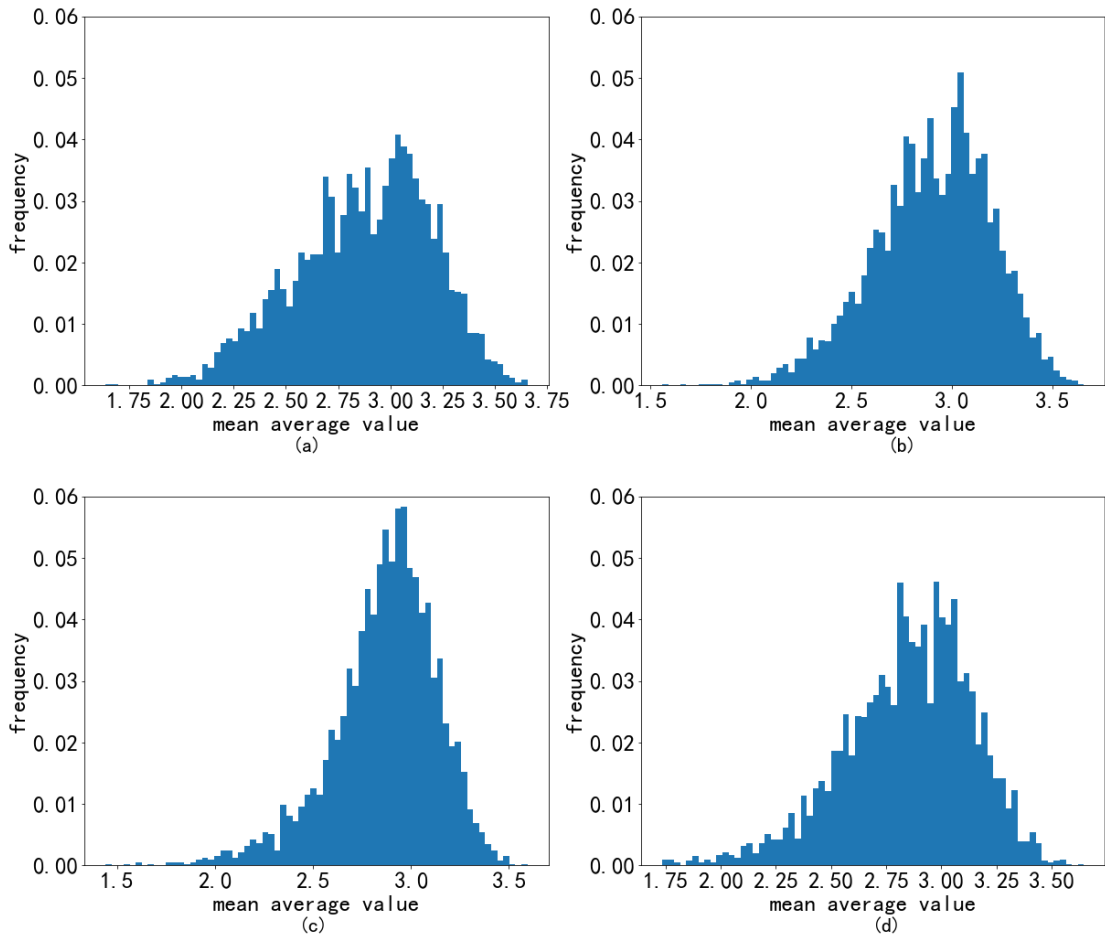


Figure 4.19: Distribution of attributes. (a) Light and color (b) Composition and depth (c) Quality (d) Semantic

level of a picture; the magnitude of 4 different aesthetic attributes spanning various level of the aesthetic appraisal (from perceptual to photographic and semantic aspects); as well as their relative importance in forming the overall aesthetic score.

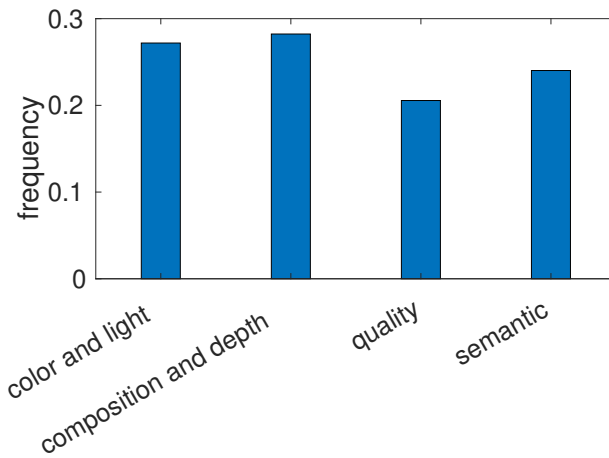


Figure 4.20: Average probability for one attribute to affect the overall aesthetic judgment.

Table 4.2: Comparison between EVA and milestone datasets ✓ means "yes", × means "not designed" or "not provided".

Dataset	scale (≈)	aesthetic label	votes/image	subject amount	subject information	train	attribute label	experiment environment
AVA [77]	255k	Distributions	78-549	×	×	×	part of challenge, semantic and style labels	photo-sharing website
AADB [55]	10k	Distributions	5-6	≈5k	ID	×	content interestingness, colour, lighting, focus and composition in 3-level rating	crowd-sourcing (AMT)
Waterloo IAA [65]	1k	mean score	26	26	general age and gender	limited	content type	lab environment
EVA	4070	Distributions	30-40	1094	personal background, demographic information, voting time	✓	assessment difficulty; aesthetic labels of perceptual, photographic and semantic aspects in 4 level rating; attribute importance.	crowd-sourcing (volunteer)

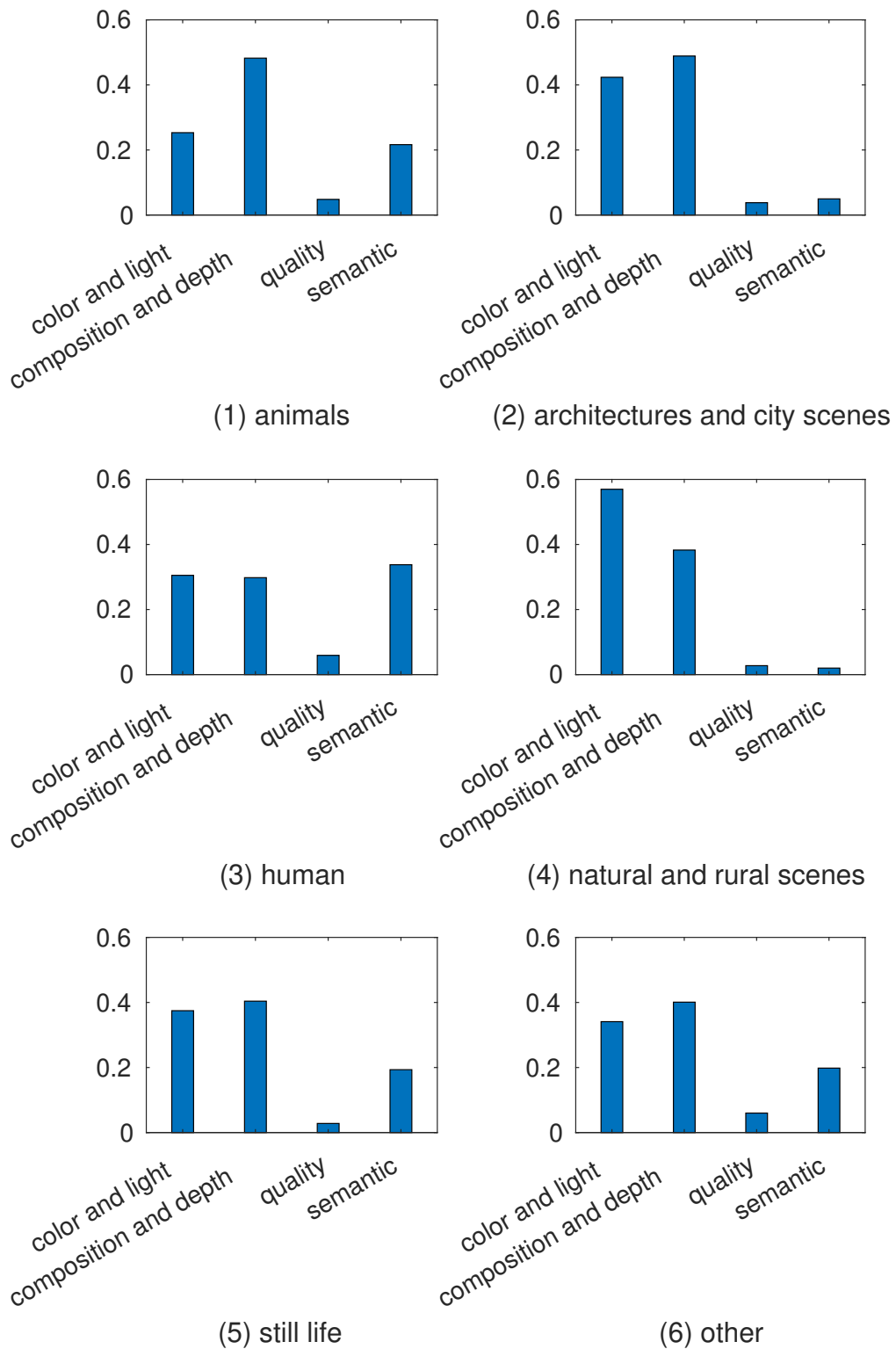


Figure 4.21: Average distribution of attributes importance per content category: (1) animals (2) architectures and city scenes (3) human (4) natural and rural scenes (5) still life (6) other

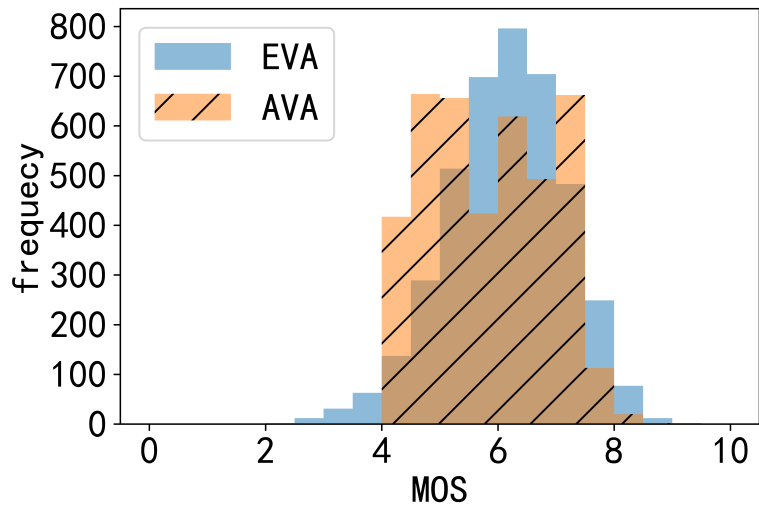


Figure 4.22: Comparison between AVA's MOS and EVA's MOS

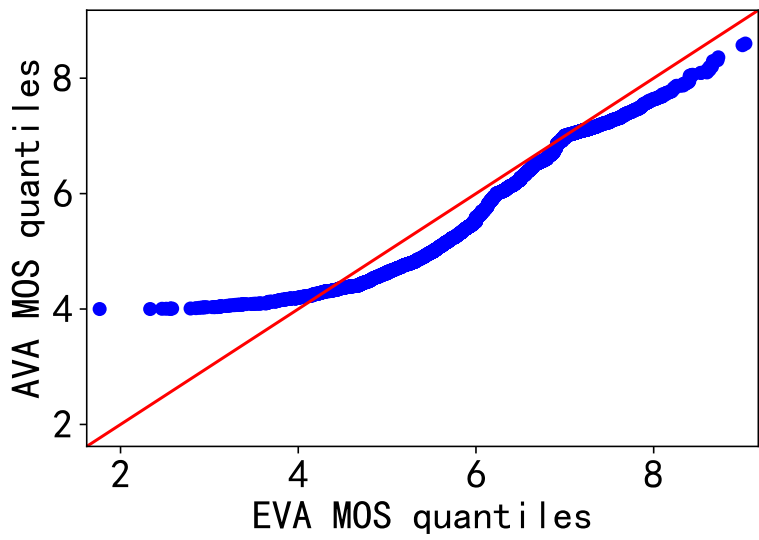


Figure 4.23: QQ-plot of AVA's MOS and EVA's MOS

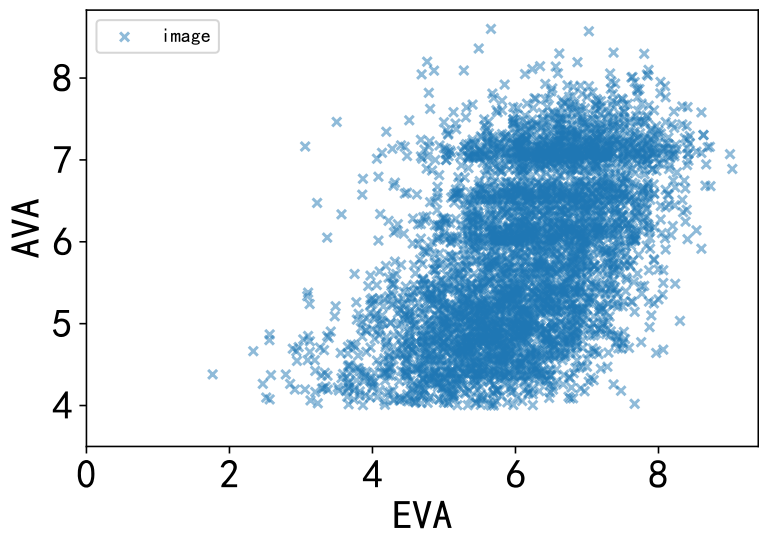


Figure 4.24: Comparison between AVA's MOS and EVA's MOS

Chapter 5

Aesthetics and Attributes

5.1 Analysis of Aesthetic Score and Attributes

To study the relation between aesthetic attributes and the overall aesthetic score, we report the Spearman's Rank-Order Correlation Coefficient (SROCC) and Pearson Linear Correlation Coefficient (PLCC) of the whole samples, as well as for each content category, in Table 5.1. The average answer of the second question in the survey, interpreted as the average personal uncertainty in an image, is denoted as "difficulty". As we used ACR, "very difficult" is recorded as 1, "difficult" is 2, "easy" is 3 and "very easy" is 4. In the third question, "very bad" is recorded as 1, "bad" is 2, "good" is 3 and "very good" is 4. Then we can get the average of votes for parameters in each image.

The correlations are divided in four groups: 1) correlation between overall aesthetic score (MOS) and average (per image) magnitude of each attribute; 2) correlation between standard deviation of the overall aesthetic score (STD) and average (per image) magnitude of each attribute; 3) correlation between difficulty and average (per image) magnitude of each attribute; and finally 4) correlation between MOS, difficulty and STD. The peak value of each row is in bold, and the one of each column is underlined.

Table 5.1: Correlation between Mean Score of Attributes and Mean Opinion Score (MOS). STD denotes the standard deviation of the global aesthetic scores per image. The numbers are SROCC/PLCC.

item	general	animal	architecture and city scenes	human	natural and rural scenes	still life	other
MOS and light and color	0.85/0.85	0.80/0.81	0.85/0.85	0.83/0.84	0.91/0.91	0.83/0.83	0.82/0.85
MOS and composition and depth	0.89/0.90	<u>0.87/0.89</u>	<u>0.88/0.89</u>	<u>0.88/0.88</u>	0.90/0.90	<u>0.88/0.89</u>	0.89/0.90
MOS and quality	0.76/0.77	0.73/0.76	0.79/0.81	0.74/0.77	0.88/0.88	0.77/0.78	0.74/0.78
MOS and semantic	0.87/0.88	0.83/0.85	0.86/0.88	0.85/0.86	0.90/0.90	0.86/0.87	0.92/0.92
STD and light and color	-0.47/-0.47	-0.36/-0.37	-0.49/-0.50	-0.43/-0.44	-0.56/-0.56	-0.41/-0.41	-0.50/-0.49
STD and composition and depth	-0.54/-0.55	-0.53/-0.51	<u>-0.52/-0.55</u>	-0.45/-0.48	-0.61/-0.61	-0.48/-0.49	-0.55/-0.53
STD and quality	-0.45/-0.45	-0.25/-0.35	-0.52/-0.52	-0.34/-0.36	-0.59/-0.59	-0.43/-0.41	-0.45/-0.42
STD and semantic	<u>-0.56/-0.58</u>	-0.47/-0.59	-0.52/-0.54	<u>-0.52/-0.56</u>	-0.59/-0.60	<u>-0.53/-0.54</u>	-0.62/-0.59
difficulty and light and color	<u>-0.62/-0.60</u>	<u>-0.58/-0.54</u>	<u>-0.60/-0.59</u>	<u>-0.56/-0.55</u>	-0.74/-0.71	<u>-0.51/-0.52</u>	<u>-0.52/-0.52</u>
difficulty and composition and depth	-0.52/-0.47	-0.50/-0.44	-0.50/-0.45	-0.44/-0.38	-0.65/-0.56	-0.42/-0.37	-0.43/-0.43
difficulty and quality	-0.49/-0.43	-0.50/-0.39	-0.47/-0.43	-0.42/-0.36	-0.67/-0.59	-0.39/-0.33	-0.40/-0.38
difficulty and semantic	-0.53/-0.48	-0.47/-0.43	-0.47/-0.45	-0.44/-0.38	-0.67/-0.59	-0.43/-0.39	-0.47/-0.46
MOS and difficulty	-0.63/-0.61	-0.60/-0.55	-0.62/-0.59	-0.57/-0.53	-0.74/-0.70	-0.53/-0.52	-0.56/-0.57
MOS and STD	-0.61/-0.62	-0.61/-0.59	-0.60/-0.62	-0.56/-0.58	-0.66/-0.66	-0.56/-0.56	-0.62/-0.58
difficulty and STD	0.24/0.24	0.16/0.15	0.22/0.22	0.16/0.15	0.37/0.35	0.13/0.13	0.24/0.24

5.1.1 Relation between Attributes and Aesthetic Score

All of the aesthetic attributes are significantly related to the general aesthetic score in a linear relationship (this is confirmed by a visual inspection of scatter plots, shown in Figure 5.1). The correlation coefficients between attributes and global score from these data seem quite similar except for the quality. This may be explained by the fact that the image quality in EVA stimuli is generally good.

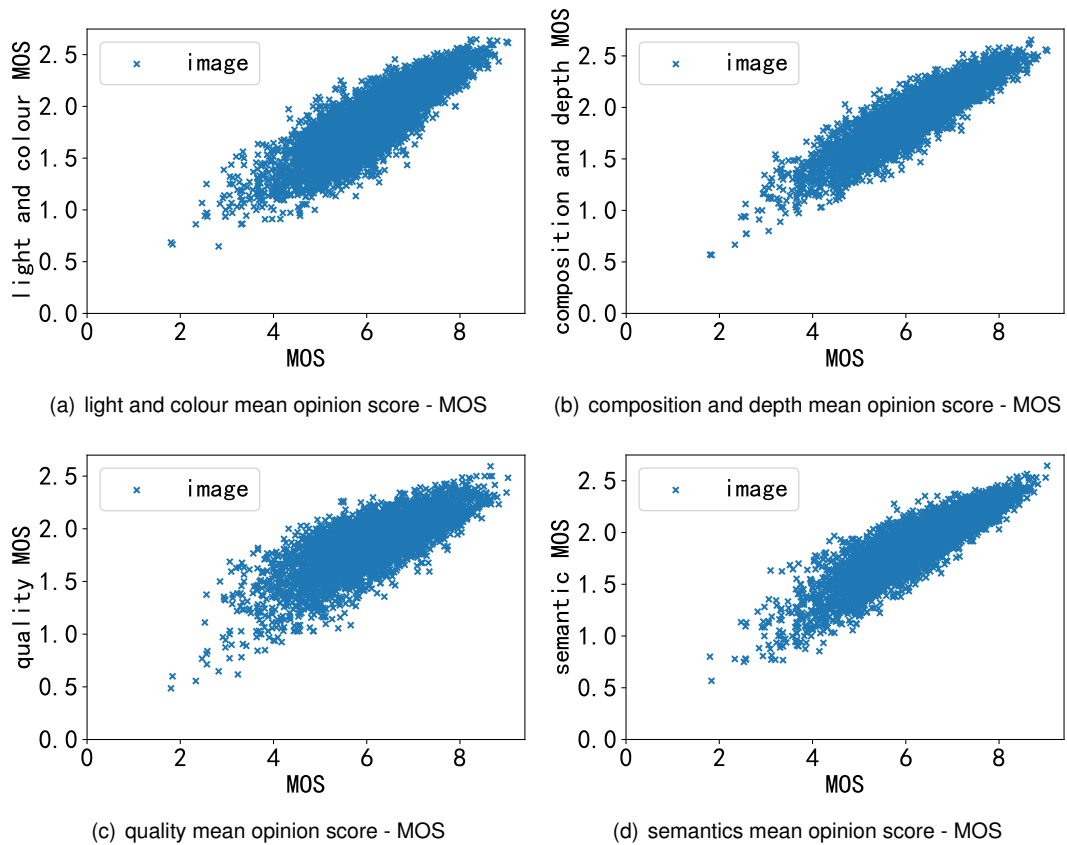


Figure 5.1: Relationship between image's MOS and each attribute's mean opinion score.

Composition and depth are the most correlated attribute with the global score in all content categories. It reaches 0.90 in PLCC for all the images. Even though in "natural and rural scenes" the most linearly related attribute is light and color, with a PLCC of 0.91, composition and depth still gets a very high correlation coefficient, with a PLCC of 0.90. Semantic preference is the most correlated attribute in the "other" category (where the content variability is higher), and it is the second most linearly related attribute among all the images. Across the categories, it can be observed that "natural and rural scenes" have a more direct relation of visual and photographic attributes to the overall score. This is probably due to the larger variety of colors, brightness, etc. than the one in a portrait or still life.

In EVA we directly elicit from observers the importance of each attribute in forming their

overall aesthetic opinion. As discussed in Section 4.2, voters had to indicate which factor(s) influenced their overall aesthetic score, among the rated attributes. This provides valuable information to explain which features of an image lead to a certain aesthetic score. Figure 4.20 reports the average probability (over all the images in the dataset) of each attribute to be selected by observers as one of those affecting the overall aesthetic score. We observe that the relative importance of each attribute is related to the correlation between the magnitude of attributes and the overall aesthetic score (Table 5.1). Quality is the least important attribute for the voters. As mentioned above, this is probably due to the fact that the selected stimuli have a relatively good image quality when displayed on personal screens and phones. Composition and depth is the most influencing attribute, and light and color are slightly less voted. Semantic is more important than quality, but it is not generally deemed the most frequent factor of explanation of the MOS by human observers.

As shown above, there is a quite good linear relationship between the average rating of an attribute for an image and the aesthetic MOS. We can then model the overall aesthetic quality of an image as a weighted sum of the quality of its attributes, that is:

$$s_i = \sum_{j=1}^4 a_j \cdot f_{ij} \quad (5.1)$$

where s_i is the MOS for image i , $a_j \geq 0$ is the weight for attribute $j \in \{1, 2, 3, 4\}$ where 1 is for color and light, 2 is composition and depth, 3 is quality, and 4 is semantic. $\sum_j a_j = 1$, and f_{ij} is the average rating of attribute j for image i . In practice, this model should include a bias term to account for the non-centered nature of the data (due to the different use of the rating scales). However, to make our analysis easier to interpret, we assume that both attributes and MOS are first normalized by removing their mean and dividing by standard deviation over the dataset.

It is possible to estimate a_j from data, by solving a constrained least-squares problem, yielding the following solution:

$$s_i \simeq 0.2877 \cdot f_{i1} + 0.2881 \cdot f_{i2} + 0.0821 \cdot f_{i3} + 0.3420 \cdot f_{i4} \quad (5.2)$$

This descriptive model fits very well our data: the root mean squared error (RMSE) of the MOS estimated by the model is 0.28, which is far below the average standard deviation of the global subjective aesthetic scores collected in the dataset (see Figure 4.17). This leads to two interesting observations about aesthetic quality assessment. First, despite its simplicity, the linearity assumption can explain effectively how aesthetics is formed. In particular, our model postulates that the weights a_j are *constant* over the dataset. This is generally not true in prac-

tice. However, even a simple zero-order approximation of these weights provides valid results: the weights in Equation (5.2) are coherent with the importance weights collected in the dataset (see Figure 4.20). Second, we conjecture that the goodness of fit of our linear model is partially due to our choice of attributes in the test design. Even if attribute scores are inter-correlated (a PCA on attribute ratings revealed that the first principal component accounts for almost 80% of the variance of the data), the well-defined nature of the attributes, which describe different qualities of the picture (from perceptual to photographic and semantic) somehow enables to easily disentangle the factors of variation of the aesthetic scores. Notice that a different selection of the attributes may have led to different models, e.g., with non-linear attribute interaction as in [2]. We believe this linear behavior is a valuable feature of EVA that might facilitate obtaining interpretable explanations of aesthetic quality.

We can also estimate importance weights from data for each image category, reported in Figure 5.2. We compare side by side the weights estimated by linear regression, with the average (normalized) importance weights collected in EVA. We observe that in general they follow a similar trend, with specific differences depending on the image category. In particular for semantics, the discrepancies are more pronounced for landscape/natural scenes and architecture, where perceptual and photographic attributes are predominant. Indeed, the impact of semantics appears to be rather complex and more difficult to describe — the definition of semantic in our dataset is quite broad and may include several co-occurring factors. Further study on this aspect is a promising research avenue for future work on aesthetic assessment.

Finally, by averaging and normalizing the binary votes over attributes, we can get a continuous, per image probability distribution of importance weights. It could then be possible to modify the linear model (5.1) to have *image-dependent* weights a_{ij} , where this time the weights are *not* computed from data, but directly obtained by eliciting them from voters. By plugging these weights into (5.1), we obtain MOS predictions with an RMSE of 0.29, just slightly worse than the global weights estimated through linear regression. This is a surprisingly good result, considering that these weights are not optimized to minimize the fitting error as in Equation (5.2). This validates the quality of the collected weights as a means to effectively explain aesthetics, and provides valuable ground-truth for future research on image aesthetics.

5.1.2 Difficulty and Subjectivity in the Aesthetic Evaluation

The results about the correlation between standard deviation and attributes' values show that the subjects' disagreement in aesthetics relates more to whether the subjects like the semantics than to the preference in low-level attributes, since the PLCC in general scores' standard deviation and semantic gets -0.58. It is similar in categories "animals", "human", "still life" and

"other". In "architecture and city scenes" and "natural and rural scene" images, composition and depth disagreement matters more than other attributes, getting -0.55 and -0.61 respectively.

Difficulty has similar correlation coefficients, but light and color is the most correlated attribute, reaching -0.60 in PLCC of all the images, and -0.71 for "natural and rural scenes" category. Difficulty and the attributes always get higher correlation in this category, and get the lowest correlation in "human" and "still life". In general, difficulty is negatively correlated with all the attributes, suggesting somehow that observers find easier to assign scores when they deem images being of high aesthetic quality. However, the small absolute values of the correlations make it difficult to draw precise conclusions at this stage.

Looking at the last group of the table, MOS has a slightly better correlation with the difficulty than STD, especially in "natural and rural scenes" category, with -0.70 and -0.66 in PLCC respectively. However, difficulty and STD have weak correlation in both SROCC and PLCC, which is 0.24 in general. This implies that average personal difficulty to judge is quite uncorrelated to group disagreement for aesthetic values [47].

5.1.3 Relation between Attributes and Subjectivity

We added the subjectivity measures "Mean Absolute Deviation around the median (MAD)", "Distance to Uniform Distribution (DUD)", and "Distance from the Maximum Entropy Distribution (MED)" following the definitions in Chapter 3. The Pearson correlation coefficient and Spearman's rank-order correlation coefficient between them and MOS, difficulty, and attributes are shown separately in Table 5.2. The biggest values of each row are in bold, and the one of each column is underlined.

For the relations between subjectivity measures and MOS, the most correlated measure is DUD in both correlations and all categories. STD and MAD show similar correlation coefficient values with MOS, around -0.6 in general; MED shows the least relationship.

For the subjectivity measures and the four attributes, they have different results across categories. MAD and STD are mostly related to semantics, and also influenced by composition and depth in some categories. DUD shows high correlation to all attributes, but relies more on semantics and composition and depth. MED nearly does not relates to any attributes.

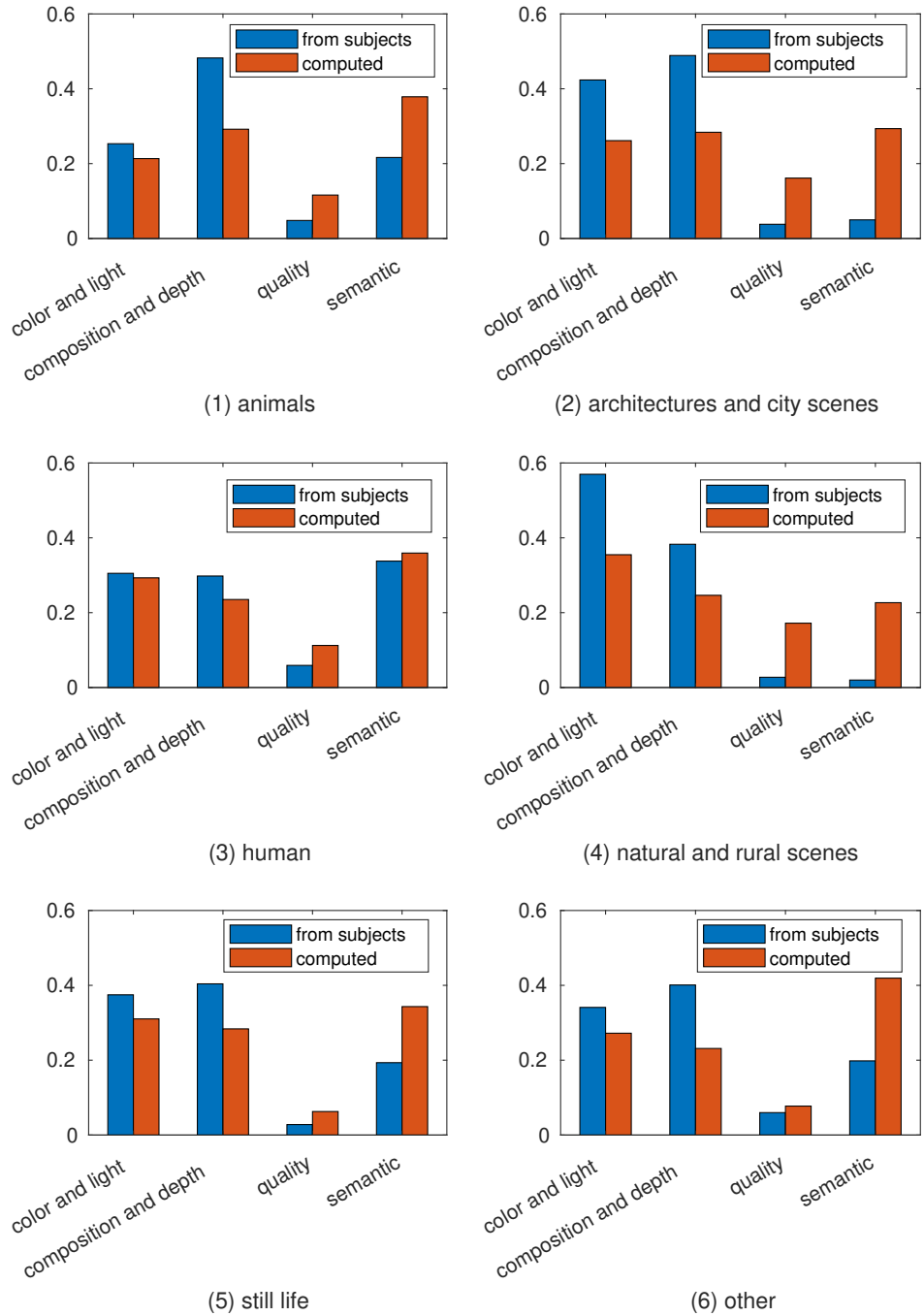


Figure 5.2: Average distribution of attributes importance per content category: (1) animals (2) architectures and city scenes (3) human (4) natural and rural scenes (5) still life (6) other

Table 5.2: Correlation between Subjectivity Measurements, Mean Opinion Score (MOS), Difficulty and Mean Score of Attributes. The numbers are SROCC/PLCC.

item	natural and rural scenes					
	general	animal	architecture and city scenes	human	still life	other
MOS and STD	-0.61/-0.62	-0.61/-0.52	-0.60/-0.62	-0.56/-0.58	-0.66/-0.66	-0.62/-0.58
MOS and MAD	-0.66/-0.65	-0.65/-0.62	-0.64/-0.64	-0.62/-0.62	-0.70/-0.70	-0.64/-0.59
MOS and DUD	<u>0.92/0.88</u>	<u>0.91/0.84</u>	<u>0.93/0.90</u>	<u>0.90/0.84</u>	<u>0.95/0.92</u>	<u>0.87/0.82</u>
MOS and MED	-0.21/-0.20	0.05/-0.01	-0.31/-0.30	-0.06/0.01	-0.46/-0.40	-0.07/-0.09
STD and light and color	-0.47/-0.47	-0.36/-0.37	-0.49/-0.50	-0.43/-0.44	-0.56/-0.56	-0.50/-0.49
STD and composition and depth	-0.54/-0.55	-0.53/-0.51	-0.52/-0.55	-0.45/-0.48	-0.61/-0.61	-0.55/-0.53
STD and quality	-0.45/-0.45	-0.25/-0.35	-0.52/-0.52	-0.34/-0.36	-0.59/-0.59	-0.45/-0.42
STD and semantic	<u>-0.56/-0.58</u>	<u>-0.47/-0.59</u>	<u>-0.52/-0.54</u>	<u>-0.52/-0.56</u>	<u>-0.59/-0.60</u>	<u>-0.62/-0.59</u>
MAD and light and color	-0.51/-0.51	-0.40/-0.40	-0.52/-0.52	-0.48/-0.49	-0.60/-0.59	-0.50/-0.48
MAD and composition and depth	-0.58/-0.58	-0.56/-0.53	-0.55/-0.57	-0.50/-0.52	-0.63/-0.64	-0.56/-0.54
MAD and quality	-0.49/-0.48	-0.39/-0.37	-0.54/-0.54	-0.39/-0.39	-0.62/-0.62	-0.44/-0.42
MAD and semantic	<u>-0.60/-0.61</u>	<u>-0.60/-0.61</u>	<u>-0.55/-0.56</u>	<u>-0.56/-0.59</u>	<u>-0.62/-0.64</u>	<u>-0.64/-0.61</u>
DUD and light and color	0.76/0.73	0.66/0.63	0.78/0.77	0.73/0.70	0.85/0.83	0.70/0.68
DUD and composition and depth	<u>0.82/0.79</u>	<u>0.79/0.74</u>	<u>0.82/0.81</u>	<u>0.76/0.73</u>	<u>0.86/0.84</u>	<u>0.78/0.73</u>
DUD and quality	0.68/0.65	0.61/0.56	0.75/0.73	0.63/0.57	0.85/0.82	0.63/0.60
DUD and semantic	<u>0.82/0.79</u>	<u>0.79/0.75</u>	<u>0.81/0.80</u>	<u>0.78/0.75</u>	<u>0.86/0.83</u>	<u>0.83/0.77</u>
MED and light and color	<u>-0.25/-0.23</u>	<u>-0.18/-0.12</u>	<u>-0.31/-0.29</u>	<u>-0.08/-0.04</u>	-0.46/-0.43	<u>-0.11/-0.12</u>
MED and composition and depth	-0.21/-0.18	-0.05/-0.00	-0.29/-0.27	-0.10/-0.04	-0.41/-0.35	-0.08/-0.08
MED and quality	-0.19/-0.15	-0.16/-0.07	-0.22/-0.20	-0.12/-0.04	-0.40/-0.35	-0.10/-0.09
MED and semantic	-0.17/-0.13	0.04/0.11	-0.29/-0.30	-0.01/0.07	-0.42/-0.35	0.00/0.01
difficulty and STD	0.24/0.24	0.16/0.15	0.22/0.22	0.16/0.15	0.37/0.35	0.24/0.24
difficulty and MAD	0.27/0.26	0.20/0.20	0.25/0.24	0.20/0.19	0.40/0.37	0.27/0.25
difficulty and DUD	<u>-0.52/-0.52</u>	<u>-0.45/-0.44</u>	<u>-0.52/-0.52</u>	<u>-0.45/-0.43</u>	-0.67/-0.63	<u>-0.44/-0.47</u>
difficulty and MED	<u>0.37/0.38</u>	<u>0.32/0.30</u>	<u>0.41/0.42</u>	<u>0.23/0.23</u>	0.53/0.52	<u>0.26/0.28</u>

Considering different categories, they do not show much difference to the overall correlations, and the most significant correlations appear in natural and rural scenes.

Difficulty shows a weak correlation to DUD, but little correlation to other measurements, which can lead to a similar result to the section 5.1.1 that difficulty and subjectivity are different concepts.

Overall, personal uncertainty is different from group disagreement, and DUD is the most correlated to mean aesthetic values.

5.2 Aesthetics and User Characteristics

In this section, we want to analyse the influence of different user's attributes based on each user's preference. Since different users voted for different images, we cannot simply use test across scores and user attributes directly. We group users by their gender, eye status, photographic level, region and age, and compute their mean score and standard deviation of votes for each user. Since Analysis of variance (ANOVA) asks for independence, normality and homogeneity of variances of the residuals, we test the normality of the user's mean general aesthetic scores and user's voting's standard deviation by Shapiro–Wilk test. Results show that at a significance level $\alpha = 0.05$, user's mean aesthetic score follows normal distribution ($p=0.120$) but standard deviation does not follow normal distribution ($p=0.002$).

We plot the distribution of each subject's voting number in the cleaned votes as Figure 4.18. In the histogram, the users who voted less than 20 images represent 32.8% of the subjects, who voted less than 25 represent 36.7%, and who voted less than 30 represent 62.9%. Subjects who voted over 1000 are limited in the histogram, representing 2.8% of the subjects. Since we want to have a limited confidence interval of these estimated means, we remove the users who voted less than 20 images or more than 1000 images. Then, we analyse the 704 subjects' mean votes and each user's standard deviation of votes.

Next, we check the conditions of groups. For the two kind of genders, male subjects represent 52% (369) and female subjects 47% (335) of the population, respectively. Standard deviation of mean scores are 1.11 for the male and 1.14 for female, and the standard deviation of standard deviation of each user's votes are not the same (0.57 for male and 0.58 for female). The standard deviation of user's mean scores in French users (0.93) and mainland Chinese users(1.15) are not the same. Other user attributes also have different standard deviation among groups, so we choose Kruskal–Wallis one-way analysis (KW test), which is the non-parametric version of ANOVA test, to check the influence of attributes to users' decisions. We compute the p-value of KW test of the variables. If it is larger than 0.05, it means that this

attribute is not significantly affecting the median value of the tested groups.

Gender

For the gender's preference, the p-value of KW test between gender and mean score is 0.246, and the one between gender and personal standard deviation is 0.166. Both of them are larger than the desired significance $\alpha = 0.05$, so we can conclude that gender does not significantly affect the median value of different groups.

We draw the QQ plot of the quantiles of female versus the quantiles of the male subjects as Figure 5.3. As can be seen, the points lie on a straight line, so their distributions almost have

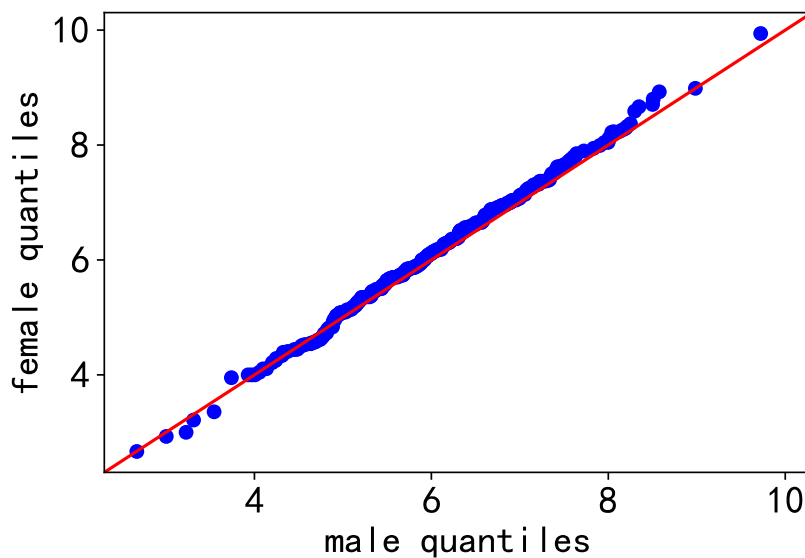


Figure 5.3: QQ plot of different genders

the same shape.

Region

The distributions of mean and standard deviations of each user's votes from France and from mainland China are normalized and shown in Figure 5.4. 537 users are from mainland China, and 79 are from France. For the mean of these groups, French users voted 5.69 with an average standard deviation of 2.08 and Chinese users voted 6.16 with a standard deviation of 1.79. This difference is also seen in Figure 5.5, that mainland Chinese users vote for higher mean score in general, but the distribution shapes are similar, as the points are aligned. According to KW test, the difference of the median of mean scores in two groups is possibly relates to the difference of regions, with a p-value approximately equals to 0.000. When looking at user's standard deviation in two groups, p-value is higher than 0.05 (6.276), so the median of user's standard deviation in two groups has a high possibility to have no difference.

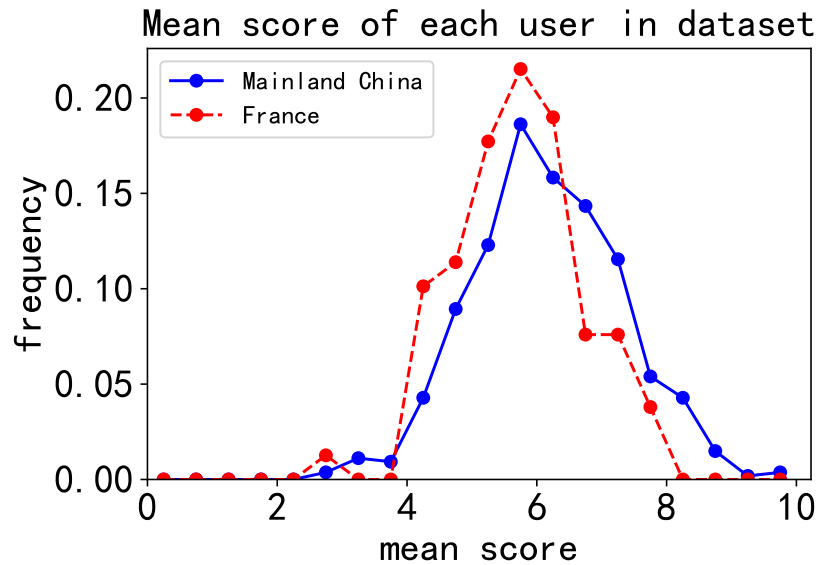


Figure 5.4: Distribution of different region's user's mean score

We did KW test on all attributes to see which attribute is influenced by subject's regions. We compute each user's mean score and standard deviation on the 4 attributes. It shows that two groups' median of user's standard deviation in all of the attributes have a high probability to be not the same, having p-values near to 0, and the median value of user's mean score in quality as well as semantics are probably not the same in two groups. This shows that region is mainly affecting high level semantics attributes and quality assessment.

Photographic level

In the users, 460 subjects are beginners, 210 subjects are intermediate level, and 34 are advanced level in photography. According to the mean of user's mean score, the mean aesthetic assessment value in each group has a small difference and decreases as the users are more trained (beginner=6.13, intermediate=5.95, advanced=5.74). KW test result shows that the mean score is affected by subject's photographic level (p-value=0.027), and the standard deviation of each subject's voting is not affected by it (p-value=0.073). This is acceptable because photography training leads to better photography knowledge, and probably make people have a different standard in aesthetics assessment.

User's mean score distributions of the untrained group (beginner level) and trained group (intermediate and advanced level) are shown in Figure 5.6. The QQ plot in Figure 5.7 shows user's mean score quantiles of them. They do not have a significant difference at the medium scale 4-7 as the points are aligned. However, the extreme mean scores are different, showing that untrained users tend to have a thinner tail to the trained voters, and is slightly skewed to the high aesthetic mean scores.

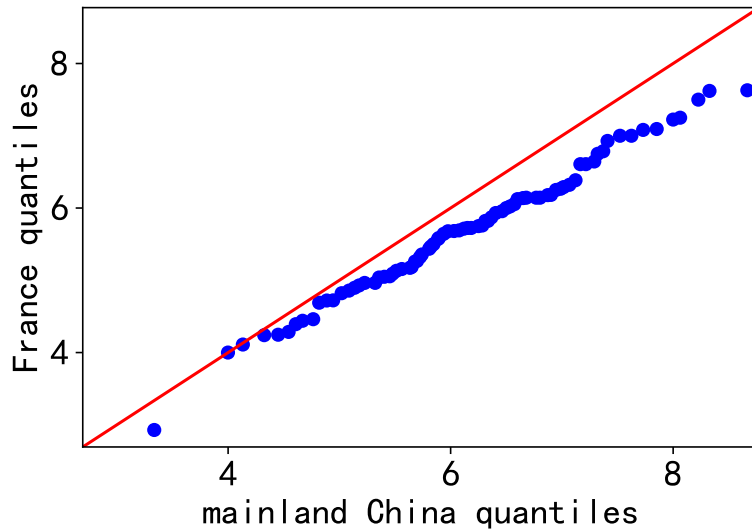


Figure 5.5: QQ plot of different regions

KW test shows that for the three groups in the four attributes, their median standard deviation have a high probability to be not the same. For light and colour, $p\text{-value}=0.017$, composition and depth and quality's $p\text{-value}=0.00$, and semantics' $p\text{-value}=0.16$. Their median mean score do not show significant difference as $p\text{-values}$ are over 0.05. This indicates that photographic knowledge and training probably affect each user's assessment range in attributes.

Thus, we compute the mean of each group's user's standard deviation on the attributes. For light and colour, beginner, intermediate and advanced users have 0.65, 0.69 and 0.72 respectively; for composition and depth, they have 0.67, 0.72 and 0.76; for quality, they have 0.62, 0.66 and 0.76; for semantics, they have 0.67, 0.70 and 0.73. The more their photographic level is, the higher mean standard deviation value the group has. This may implies that trained people are more sensitive to evaluate attributes.

Above all, we can conclude that photography level significantly affects user's general aesthetic assessment, but has limited influence in attributes' values.

Eye status

As for the eye status, lab based quality assessment subjective tests usually hire subjects without colourblindness, and they record if the subjects are wearing glasses, while crowd-sourcing method usually ignore this factor.

We have 442 subjects wearing glasses, 14 subjects are colourblind but without glasses, 8 subjects are colourblind and wearing glasses, and 240 subjects do not have these conditions.

We first check if wearing glasses significantly affect the median of user's general aesthetic assessment, and the KW test result shows that all $p\text{-values}$ are over 0.05, showing the little

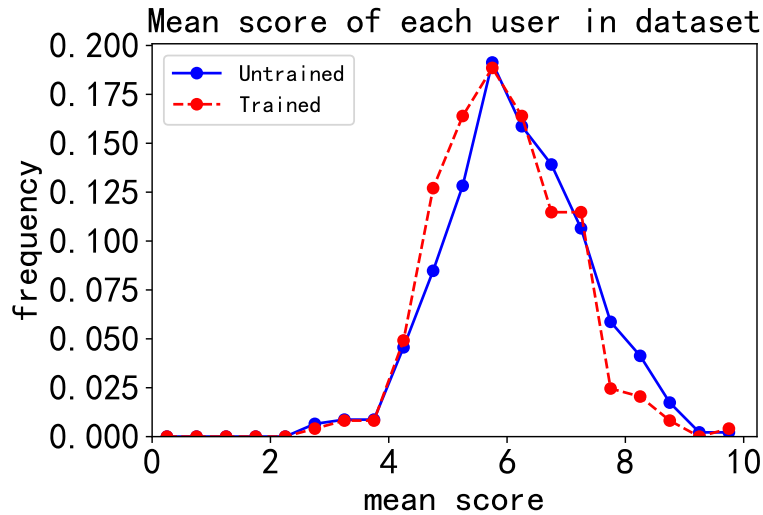


Figure 5.6: Distribution of different photographic level's user's mean score

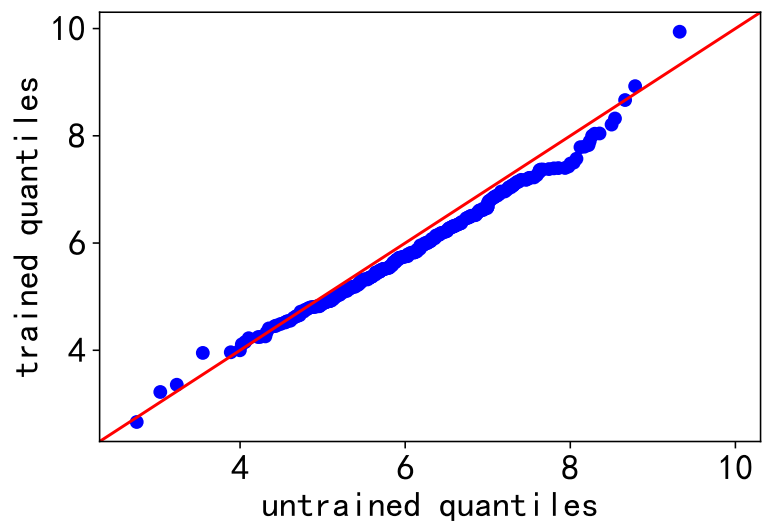


Figure 5.7: QQ plot of beginners and intermediate subjects

influence.

Then, we plot Figure 5.8 and Figure 5.9 to see the influence of colour blindness. They have different distribution. For general scores, the medians of user's standard deviation in colour blinded group and normal vision group have a high probability to be different, with $p\text{-value}=0.039$, while the mean score has $p\text{-value}=0.689$. This shows the colour vision condition is not affecting the general aesthetic scores obviously.

To test if colour vision affects the attributes, especially visual attributes, we applied KW test on user's mean attribute scores in our dataset. We can see from Figure 5.10 that the shape of distribution of user's mean attribute score changes with colourblind. However, the KW test result shows that different eye condition groups has a $p\text{-value}=0.513$ in user's mean colour and light score and $p\text{-value}=0.016$ in user's light and colour's standard deviation, which

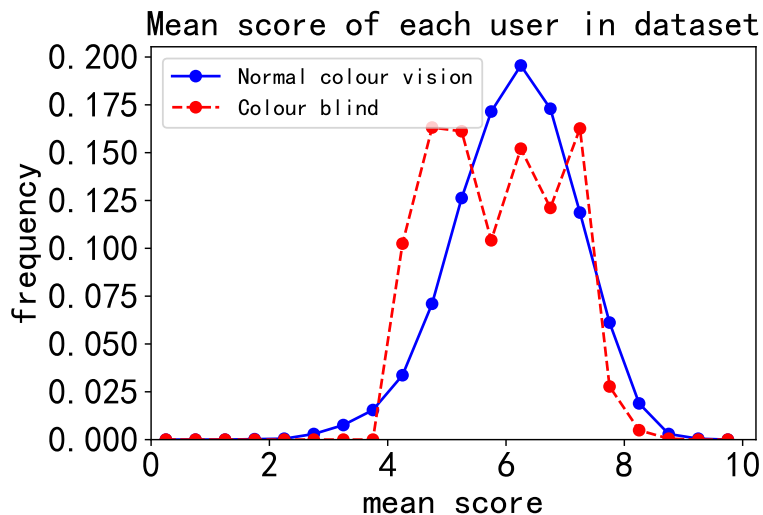


Figure 5.8: Distribution of different eye status' user's mean score

shows the median of standard deviation is significantly affected by the colour vision condition. It may be because our attribute definition contains factors more than colour harmony, so the colour blinded can vote as well as others. Also, it may be because colour blinded volunteer numbers are limited.

Age

We grouped the users by age (at 2020) to be three groups: 445 users are aged 18-30; 123 users are aged 31-50; 135 users are over 51. The distributions are illustrated in Figure 5.11. The average of each group's user's mean score are 5.96, 6.27 and 6.17 respectively. From Figure 5.12, we can see that subjects aged 18-30 have similar quantiles with 31-50 with mean score before 6, but their quantiles are not aligned afterwards. As the age gets bigger, the distributions get more right skewed. KW test shows that the median of each group's user's mean score is significantly affected by age, as $p\text{-value}=0.041$.

Then, we check different age group's preference in attributes. The median of different group's mean light and colour as well as quality have a big possibility to be different, with $p\text{-value}$ equals to 0.035 and 0.045 respectively. This shows age affects the visual attributes more than other attributes.

5.3 Conclusion

In this chapter, we analysed the influencing attributes to image aesthetics among the collected EVA data.

Statistical analysis on the collected data shows that the chosen attributes are linearly related

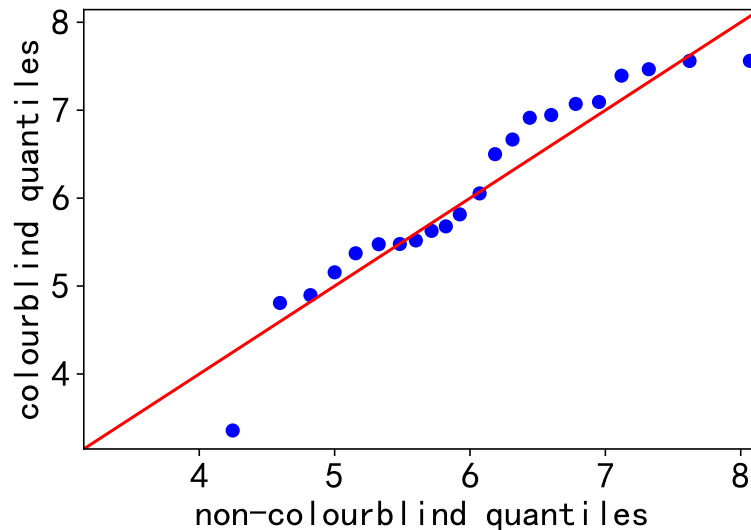


Figure 5.9: QQ plot of different eye status' user's mean score. Non-colourblind includes normal subjects and the ones with glasses; colourblind includes only having colourblindness and having colourblindness and glasses at the same time.

to the overall aesthetic score. This leads to proposing a simple, yet effective, linear model to explain aesthetic score formation. We find that the subjective importance weights expressed by observers provide a surprisingly good fit to data under this model, which demonstrates the goodness of the collected dataset. In particular, EVA enables to estimate the importance of each aesthetic factor *per image*, thus effectively enabling the explanation of aesthetic scores.

We also found that mean aesthetic personal difficulty of an image has little correlation to subjectivity measures. Subjectivity measures show different relations to mean opinion score and difficulty.

The Kruskal–Wallis one-way analysis of variance test on different group of users show that geographical region, photographic level and age significantly affect subject's mean image aesthetic votes, standard deviation of the votes, and a few aesthetic attributes assessment. Eye status is not significantly affecting aesthetics and visual attributes.

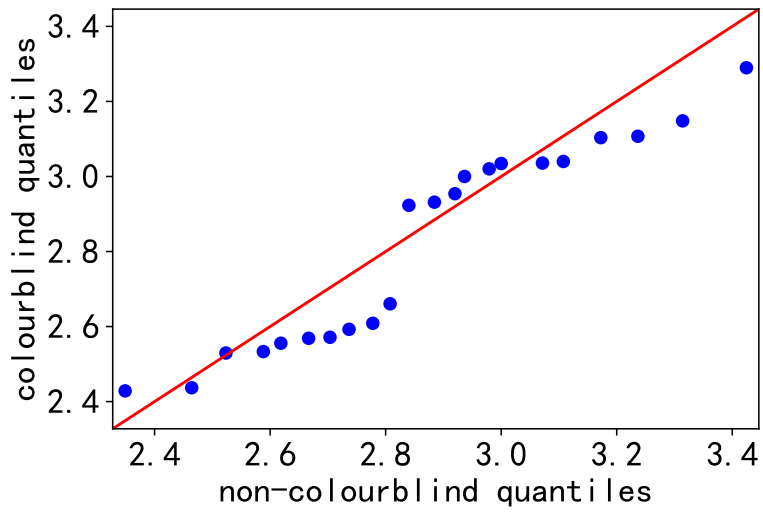


Figure 5.10: QQ plot of different eye status' user's mean colour and light score

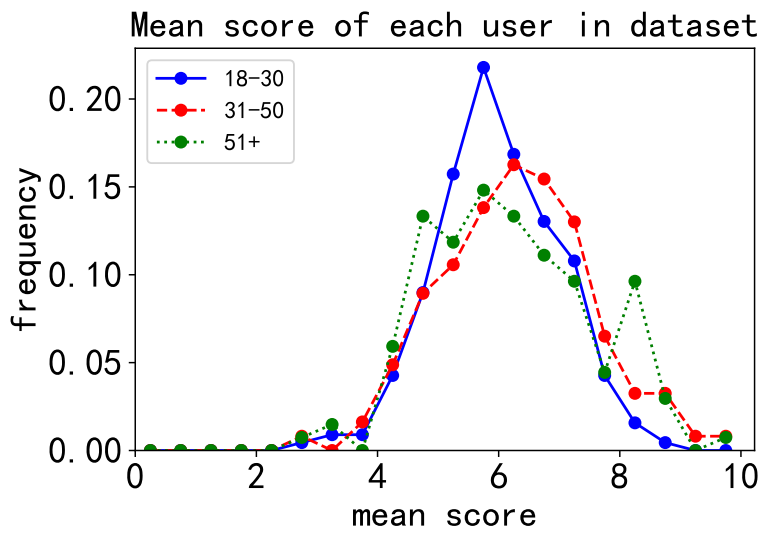


Figure 5.11: Distribution of different ages' user's mean votes

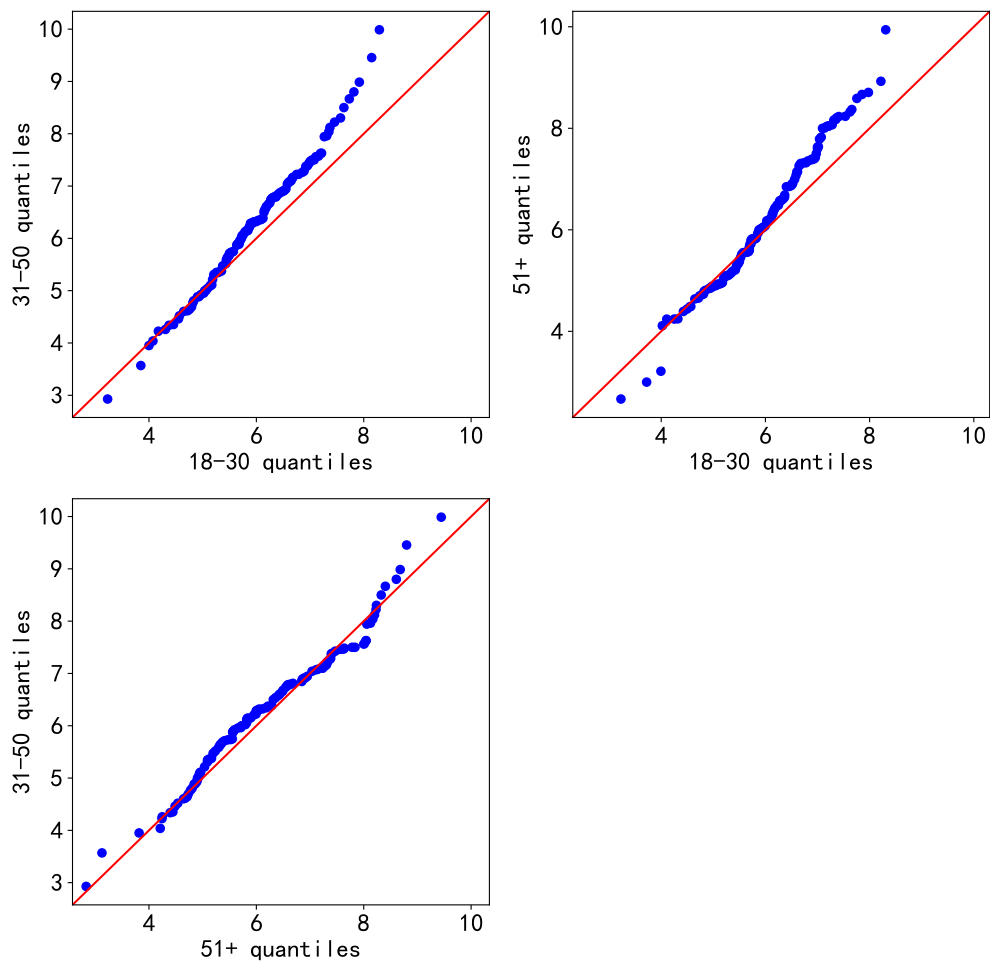


Figure 5.12: QQ plots of different ages' user's mean votes

Chapter 6

Conclusion and Perspective

6.1 Summary of the Contributions

In this thesis, we first studied about subjectivity, and then provided data and analysis towards computing explainable aesthetic quality assessment. In Chapter 2, we went through the development of image aesthetic quality prediction. Then, we introduced the state-of-the-art of aesthetic subjectivity. Furthermore, we reviewed existing image aesthetic datasets, indicating that the lack of well-labelled datasets limits the exploration of subjectivity and interpretability for image aesthetics.

In Chapter 3, we defined four aesthetic score subjectivity measurements for describing the degree of consensus about an image's aesthetic value. They are inspired by statistics and information theoretical principles, and we predicted them with deep neural networks. At the same time, we compared two possible subjectivity prediction frameworks, and we found that computing subjectivity from ground-truth distribution and predicting it directly has a better performance than computing the subjectivity from predicted score distributions. This shows that predicting only the aesthetic score distribution is not enough for predicting the subjectivity of image aesthetics. We also did the experiment in improving the prediction of aesthetic mean scores by the help of aesthetic subjectivity measurements. Although it has limited improvement, information theory inspired subjectivity measurement performs better than the statistical measurements in general in the AVA dataset.

In Chapter 4, we proposed a new Explainable Aesthetic Dataset (EVA) where aesthetics and attribute labels are annotated totally by human. It consists of 4070 images with 30-40 votes covering general aesthetic score, difficulty, 4 attributes, attribute importance and user information. We described the design of the survey and experiment method, the choice of images and users, data collection and the summary of information. This shows that EVA has

many advantages comparing to the existing datasets, especially it overcomes the noisy labels caused by ambiguous task description, limited subjects and lack of well human-labelled difficulty and attributes.

In Chapter 5, we analysed influencing attributes based on EVA dataset. We found that the chosen attributes have a linear relationship with the general aesthetic score. Difficulty is not the same with subjectivity descriptors. In particular, EVA enables to estimate the importance of each aesthetic factor per image, thus effectively enabling the explanation of aesthetic scores. The Kruskal–Wallis one-way analysis of variance test on demographic/cultural background with other characteristics of the observers show that geographical region, photographic level and age are significantly affecting user’s average aesthetic assessment from general score and attributes.

6.2 Perspectives

Following our test of subjectivity measurements, new dataset and analysis towards computing explainable aesthetic predictions, further experiments can be done.

More subjectivity measurements can be proposed and compared to our measurements. Sophisticated algorithms can be used to compare the performance of subjectivity measurements, and to improve the prediction of distribution or mean aesthetic score. For example, we can add the subjectivity as a penalty of the loss function of predicting distributions. By having a better accuracy, it can be used in applications like retrieval, where images with high aesthetic values with a bigger consensus can rank higher in the results.

Since our proposed EVA dataset is available online, it can be expanded under the same condition of experiment. Future efforts can be made in recruiting a larger variety of subjects or exploring new ways to clean the data based on other collected information. Also, some data control method for crowd-sourcing can be borrowed from other field, e.g., the agreement multigraph to check user’s agreement in [118]; we may learn from the reliability screening mechanisms in [39], where authors found that experts choose same answers for different images less frequent than non-expert subjects.

More analysis inside EVA dataset can be done. Since we have user’s choice on factor importance, we can learn the preference of a group of users. Also, by analysing attributes and factor importance in each user group, we can know more about how does the user background influence image aesthetics. Previous research believes the standard deviation of a score histogram can show the difficulty of image’s aesthetic assessment [43], while our analysis in EVA shows little relation between the personal difficulty and the group’s standard deviation. There-

fore, how to know how difficult an image is in aesthetic assessment still needs to be studied. What is more, personalised image aesthetics study can be done on EVA since it records each user's preference.

Appendix A

Aesthetics in Media Memorability Prediction

This appendix¹ is the work of using image aesthetic assessment on videos to improve video memorability prediction.

In this appendix we present the contribution and results of the participation of the UPB-L2S team to the MediaEval 2019 Predicting Media Memorability Task. The task requires participants to develop machine learning systems able to predict automatically whether a video will be memorable for the viewer, and for how long (e.g., hours, or days). To solve the task, we investigated several aesthetics and action recognition-based deep neural networks, either by fine-tuning models or by using them as pre-trained feature extractors. Results from different systems were aggregated in various fusion schemes. Experimental results are positive showing the potential of transfer learning for this tasks.

A.1 Introduction

Media Memorability was studied extensively in recent years, playing an important role in the analysis of human perception and understanding of media content. This domain was approached by numerous scientists from different perspectives and fields of study, including psychology [94, 7] and computer vision [90, 16], while several works analyzed the correlation between memorability and other visual perception concepts like interestingness and aesthetics [41, 19]. In this context, the MediaEval 2019 Predicting Media Memorability task requires participants to create systems that can predict the short-term and long-term memorability of a set of soundless videos. The dataset, annotation protocol, precomputed features, and ground

¹This work has been published in *MediaEval 2019 Workshop*, Oct 2019, Sophia Antipolis, France.

truth data are described in the task overview paper [18].

A.2 Approach

For our approach, we used several deep neural network models based on image aesthetics and action recognition. For the first category, we fine-tuned the aesthetic deep model presented in [47]. It is based on the ResNet-101 architecture [37]. For the action recognition networks, we used features extracted from the I3D [8] and TSN [109] networks and attempted to augment these features with the C3D [106] features provided by the task organizers. Finally, we performed some late fusion experiments to further improve the results of these individual runs. Figure A.1 summarizes and presents these approaches. The approaches are detailed in the following.

A.2.1 Aesthetics networks

The aesthetic-based approach modifies the ResNet-101 architecture [37], trained on the AVA dataset [77] for the prediction of image aesthetic value, following the approach described in [47]. This approach generates a deep neural model that can process single image aesthetics and must be fine-tuned to process the short and long term memorability of videos. To generate a training dataset that will support the fine-tuning process, we extracted key-frames in two ways: (i) key frames from the 4th, 5th, and 6th second of each sample; (ii) one key frame every two seconds to test multi-frame training. In the retraining stage of the network for the memorability task, the provided devset is randomly split into three parts, with 65% of the samples representing the training set, 25% the test set and 10% the validation set. We adapted the last layer for this task by creating a fully connected layer with 2,048 inputs and 1 output. During the fine-tuning process, we applied mean square error as loss function, using an initial learning rate of 0.0001. We ran the training process for 15 epochs, with a batch size of 32.

A.2.2 Action recognition networks

Apart from the precomputed C3D features, we extracted the "Mixed_5" layer from the I3D network [8], trained on the Kinetics dataset [52] and the "Inception_5" layer of the TSN network [109], trained on the UCF101 dataset [97]. These features were used as inputs for a Support Vector Regression algorithm that generates the final memorability scores. We conducted preliminary early fusion tests with combinations of these features in order to select the best possible combinations, testing both each feature vector individually and all possible com-

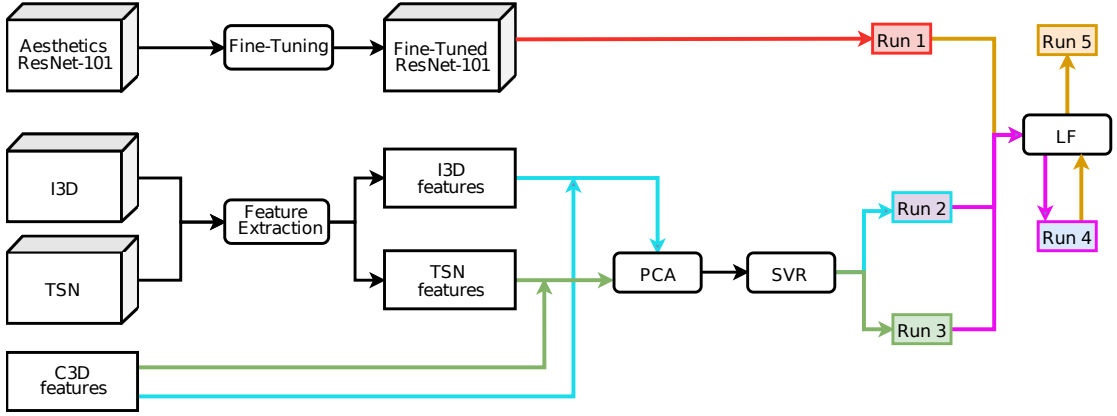


Figure A.1: The diagram of the proposed solution.

binations of two feature vectors. We also employed a PCA dimensionality reduction, reducing the size of each vector to 128 elements. Finally, to train the SVR system, we used a random 4-fold approach, with 75% of the data representing the training set and 25% representing the validation set. We used parameter tuning for the SVR model, via a RBF kernel and performing a grid search with two parameters: the C parameter and the gamma parameter (taking values 10^k , where $k \in [-4, \dots, 4]$).

Table A.1: Results of the proposed runs (preliminary experiments on *devset*, and official results on *testset*).

Run	System description	Devset - Spearman's ρ		Testset - Spearman's ρ	
		Short-term	Long-term	Short-term	Long-term
run1	Aesthetic-based	0.448	0.230	0.401	0.203
run2	Action-based (TSN+I3D)	0.473	0.259	0.45	0.228
run3	Action-based (C3D+I3D)	0.433	0.204	0.386	0.184
run4	Late Fusion Action-based (run2 + run3)	0.466	0.200	0.439	0.218
run5	Late Fusion Aesthetic and Action (run1 + run2)	0.494	0.265	0.477	0.232

A.2.3 Late fusion

We employed several late fusion schemes on the best performing systems, trying to benefit from their combined strengths. We used three different strategies for combining these scores, namely: (i) LFMax, where we took the maximum score for each media sample; (ii) LFMIn, where we took the minimum score; (iii) LFWeight, where each score from different samples was multiplied with a weight w . We assigned each weight varying values according to the formula $w = 1 - r/c$, where the rank r had the value 0 for the best performing system, 1 for the second best and so on, and c represents a coefficient that dictates rank influence on the weights.

A.3 Experimental results

The development dataset consists of 8,000 videos, annotated with short and long term memory scores, while the test dataset consists of 2,000 videos. The official metric used in the task is Spearman's rank correlation (ρ). The best performing systems in the development phase are selected, retrained on the whole devset by using the optimal parameters and lastly run on the testset data.

A.3.1 Results on the devset

During the tests performed on the devset, several systems and combinations of parameters stood out as best performers. Table A.1 shows the performances recorded by the best performing aesthetic, action-based, and late fusion systems.

We used several dataset variations in retraining the aesthetic-based deep network. More precisely, we found that, for the short-term memorability, the best performing systems were the ones trained with keyframes extracted from the 5th second and the ones extracted from the multi-frame approach. The results were both similar with a Spearman's ρ of 0.45. On the other hand, in the long-term memorability subtask we found that the best performing systems were the ones trained with keyframes from the 5th frame. Although this may seem somewhat surprising, giving that bigger data sets usually account for better results, we believe that the reason behind this is that each video contains only one scene. Therefore not much additional information is given to the system when more frames are extracted because the frames are very similar. However, we would also like to point out that the results for the other frame extraction schemes were not much lower than these.

Regarding the 3D action-recognition based systems, we noticed that individual systems, based on only one feature vector (TSN, I3D or C3D) had a low performance, with a Spearman's ρ score of under 0.42. This performance further dropped when we used the original vectors, without applying PCA reduction, therefore demonstrating the positive influence that dimensionality reduction has on the final results. Therefore we decided to apply an early fusion scheme, where we tested all the possible combinations of the feature vectors, by concatenating them. The best performing combinations were TSN + I3D and C3D + I3D.

Finally, in the late fusion part of the experiment, we generally decided to test late fusion schemes between the two action-recognition based systems and between the best performing action-recognition system (TSN + I3D) and the aesthetic-based system. In general, results for the LFMin systems were underperforming, while the LFMax systems were better than their components, but without bringing a significant increase in results. The best performing late

fusion schemes proved to be based on LFWeight, more precisely using a c value of 5. This was an expected result, as it confirms some of our previous work in other MediaEval tasks [17].

A.3.2 Results on the testset

For the final phase, we retrained all the systems on the entire set of videos from devset, using the parameters computed in the previous phases and tested them on the videos from the testset. Table A.1 presents also the results for this phase.

As expected, the best performance comes from a late fusion system using both aesthetic and action-based components (short-term $\rho = 0.477$ and long-term $\rho = 0.232$). Generally, we observe that the system ranking for the submitted systems is consistent with the one we observed during the development phase, however, the results are lower than those predicted then, with significant drops in performance for the aesthetic-based system and the action-based (C3D + I3D) approaches. In terms of single-system performance, the action-based TSN + I3D system performs best, followed by the aesthetic-based system.

A.4 Conclusions

In this appendix we presented the UPB-L2S approach for predicting media memorability at MediaEval. We created a framework that uses aesthetic and action recognition based systems and some late fusion combinations of these systems, that predict short-term and long-term memorability scores for soundless video samples. The results show that these systems are able to individually predict these scores, while the best results are achieved via late fusion weighted schemes. This enforces the idea of better exploiting transfer learning to tasks where labeled data are in particular hard to obtain.

Appendix B

Author's publications during 3 years of PhD

Papers published in international conferences :

1. **C. Kang**, G. Valenzise, and F. Dufaux. "Predicting Subjectivity in Image Aesthetics Assessment." 2019 IEEE *21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019.
2. **C. Kang**, G. Valenzise, and F. Dufaux. "EVA: An Explainable Visual Aesthetics Dataset." *Joint Workshop on Aesthetic and Technical Quality Assessment of Multimedia and Media Analytics for Societal Trends, ACM Multimedia Conference*. ACM, 2020.
3. MG. Constantin, **C. Kang**, G. Dinu, F. Dufaux, G.Valenzise, and B. Ionescu. "Using Aesthetics and Action Recognition-based Networks for the Prediction of Media Memorability." *MediaEval Benchmarking Initiative for Multimedia Evaluation (MediaEval) Workshop*. 2019.

Bibliography

- [1] Aesthetics. [EB/OL]. <https://en.wikipedia.org/wiki/Aesthetics> Accessed June 4, 2020.
- [2] T. O. Aydin, A. Smolic, and M. Gross. Automated aesthetic analysis of photographic images. *IEEE transactions on visualization and computer graphics*, 21(1):31–42, 2014.
- [3] R. Barthes. *La chambre claire*. Gallimard Paris, 1980.
- [4] R. Barthes. The photographic message. *Theorizing communication: readings across traditions*, pages 191–199, 2000.
- [5] S. Bianco, R. Cadene, L. Celona, and P. Napoletano. Benchmark analysis of representative deep neural network architectures. *IEEE Access*, 6:64270–64277, 2018.
- [6] S. Bianco, L. Celona, and R. Schettini. Aesthetics assessment of images containing faces. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2820–2824. IEEE, 2018.
- [7] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008.
- [8] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [9] M. Chambe, R. Cozot, and O. Le Meur. Behaviour of recent aesthetics assessment models with professional photography. 2019.
- [10] H. Chang, F. Yu, J. Wang, D. Ashley, and A. Finkelstein. Automatic triage for a photo series. *ACM Transactions on Graphics (TOG)*, 35(4):148, 2016.

- [11] K.-Y. Chang, K.-H. Lu, and C.-S. Chen. Aesthetic critiques generation for photos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3514–3523, 2017.
- [12] A. Chatterjee and O. Vartanian. Neuroscience of aesthetics. *Annals of the New York Academy of Sciences*, 1369(1):172–194, 2016.
- [13] Y.-L. Chen, J. Klopp, M. Sun, S.-Y. Chien, and K.-L. Ma. Learning to compose with professional photographs on the web. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 37–45, 2017.
- [14] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [15] W.-T. Chu, Y.-K. Chen, and K.-T. Chen. Size does matter: How image size affects aesthetic perception? In *Proceedings of the 21st ACM international conference on Multimedia*, pages 53–62, 2013.
- [16] Cohendet, Romain and Demarty, Claire-Hélène, and Duong, Ngoc Q. K. and Engilberge, Martin. Videomem: Constructing, analyzing, predicting short-term and long-term video memorability. In *International Conference on Computer Vision (ICCV)*, 2019.
- [17] M. G. Constantin, B. A. Boteanu, and B. Ionescu. Lapi at mediaeval 2017-predicting media interestingness. In *MediaEval*, 2017.
- [18] M. G. Constantin, B. Ionescu, C.-H. Demarty, N. Q. K. Duong, X. Alameda-Pineda, and M. Sjöberg. Predicting media memorability task at mediaeval 2019. In *Proc. of MediaEval 2019 Workshop, Sophia Antipolis, France, Oct. 27-29, 2019*, 2019.
- [19] M. G. Constantin, M. Redi, G. Zen, and B. Ionescu. Computational understanding of visual interestingness beyond semantics: literature survey and analysis of covariates. *ACM Computing Surveys (CSUR)*, 52(2):25, 2019.
- [20] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [21] C. Cui, H. Liu, T. Lian, L. Nie, L. Zhu, and Y. Yin. Distribution-oriented aesthetics assessment with semantic-aware hybrid network. *IEEE Transactions on Multimedia*, 21(5): 1209–1220, 2018.
- [22] C. Cui, W. Yang, C. Shi, M. Wang, X. Nie, and Y. Yin. Personalized image quality assessment with social-sensed aesthetic preference. *Information Sciences*, 512:780–794, 2020.

- [23] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision*, pages 288–301. Springer, 2006.
- [24] R. Datta, J. Li, and J. Z. Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *2008 15th IEEE International Conference on Image Processing*, pages 105–108. IEEE, 2008.
- [25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.
- [26] Y. Deng, C. C. Loy, and X. Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017.
- [27] Y. Deng, C. C. Loy, and X. Tang. Aesthetic-driven image enhancement by adversarial learning. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 870–878, Seoul, Korea, 2018. ACM.
- [28] Z. Dong and X. Tian. Effective and efficient photo quality assessment. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2859–2864. IEEE, 2014.
- [29] Z. Dong, X. Shen, H. Li, and X. Tian. Photo quality assessment with dcnn that understands image well. In *International Conference on Multimedia Modeling*, pages 524–535. Springer, 2015.
- [30] F. Farhat, M. M. Kamani, and J. Z. Wang. Captain: Comprehensive composition assistance for photo taking. *arXiv preprint arXiv:1811.04184*, 2018.
- [31] F. Gao, Z. Li, J. Yu, J. Yu, Q. Huang, and Q. Tian. Style-adaptive photo aesthetic rating via convolutional neural networks and multi-task learning. *Neurocomputing*, 395:247–254, 2020.
- [32] M. Gheisari, G. Wang, and M. Z. A. Bhuiyan. A survey on deep learning in big data. In *2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC)*, volume 2, pages 173–180. IEEE, 2017.
- [33] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

- [34] K. A. Hallgren. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23, 2012.
- [35] J. He, L. Wang, W. Zhou, H. Zhang, X. Cui, and Y. Guo. Viewpoint assessment and recommendation for photographing architectures. *IEEE transactions on visualization and computer graphics*, 2018.
- [36] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.
- [37] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [38] T. Hoßfeld, R. Schatz, and S. Egger. Sos: The mos is not enough! In *2011 third international workshop on quality of multimedia experience*, pages 131–136. IEEE, 2011.
- [39] V. Hosu, H. Lin, and D. Saupe. Expertise screening in crowdsourcing image quality. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2018.
- [40] T. Ishizu and S. Zeki. The brain’s specialized systems for aesthetic and perceptual judgment. *European Journal of Neuroscience*, 37(9):1413–1420, 2013.
- [41] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1469–1482, 2013.
- [42] W. Jiang, A. C. Loui, and C. D. Cerosaletti. Automatic aesthetic value assessment in photographic images. In *2010 IEEE International Conference on Multimedia and Expo*, pages 920–925. IEEE, 2010.
- [43] B. Jin, M. V. O. Segovia, and S. Süssstrunk. Image aesthetic predictors based on weighted CNNs. In *IEEE International Conference on Image Processing*, pages 2291–2295, Phoenix, AZ, USA, October 2016. IEEE.
- [44] X. Jin, L. Wu, X. Li, S. Chen, S. Peng, J. Chi, S. Ge, C. Song, and G. Zhao. Predicting aesthetic score distribution through cumulative jensen-shannon divergence. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

- [45] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5):94–115, 2011.
- [46] M. Kairanbay, J. See, and L.-K. Wong. Towards demographic-based photographic aesthetics prediction for portraits. In *International Conference on Multimedia Modeling*, pages 531–543, Bangkok, Thailand, 2018. Springer.
- [47] C. Kang, G. Valenzise, and F. Dufaux. Predicting subjectivity in image aesthetics assessment. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, Kuala Lumpur, Malaysia, 2019. IEEE.
- [48] Y. Kao, C. Wang, and K. Huang. Visual aesthetic quality assessment with a regression model. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1583–1587. IEEE, 2015.
- [49] Y. Kao, K. Huang, and S. Maybank. Hierarchical aesthetic quality assessment using deep convolutional neural networks. *Signal Processing: Image Communication*, 47:500–510, 2016.
- [50] Y. Kao, R. He, and K. Huang. Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing*, 26(3):1482–1495, 2017.
- [51] S. M. Kay. *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993.
- [52] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- [53] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 419–426. IEEE, 2006.
- [54] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [55] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, pages 662–679, Amsterdam, Netherlands, 2016. Springer.
- [56] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

- [57] M. Kucer, A. C. Loui, and D. W. Messinger. Leveraging expert feature knowledge for predicting image aesthetics. *IEEE Transactions on Image Processing*, 27(10):5100–5112, 2018.
- [58] D. Kuzovkin, T. Pouli, R. Cozot, O. L. Meur, J. Kerverc, and K. Bouatouch. Context-aware clustering and assessment of photo collections. In *Proceedings of the symposium on Computational Aesthetics*, pages 1–10, 2017.
- [59] D. Kuzovkin, T. Pouli, R. Cozot, O. Le Meur, J. Kerverc, and K. Bouatouch. Image selection in photo albums. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 397–404, 2018.
- [60] H.-J. Lee, K.-S. Hong, H. Kang, and S. Lee. Photo aesthetics analysis via dcnv feature encoding. *IEEE Transactions on Multimedia*, 19(8):1921–1932, 2017.
- [61] J.-T. Lee and C.-S. Kim. Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1191–1200, 2019.
- [62] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436, 2018.
- [63] D. Li, H. Wu, J. Zhang, and K. Huang. A2-rl: Aesthetics aware reinforcement learning for image cropping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8193–8201, 2018.
- [64] D. Li, H. Wu, J. Zhang, and K. Huang. Fast a3rl: Aesthetics-aware adversarial reinforcement learning for image cropping. *IEEE Transactions on Image Processing*, 28(10):5105–5120, 2019.
- [65] W. Liu and Z. Wang. A database for perceptual evaluation of image aesthetics. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1317–1321, Beijing, China, 2017. IEEE.
- [66] Z. Liu, Z. Wang, Y. Yao, L. Zhang, and L. Shao. Deep active learning with contaminated tags for image aesthetics assessment. *IEEE Transactions on Image Processing*, 2018.
- [67] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 457–466. ACM, 2014.

- [68] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 17(11):2021–2034, 2015.
- [69] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 990–998, 2015.
- [70] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *2011 International Conference on Computer Vision*, pages 2206–2213. IEEE, 2011.
- [71] S. Ma, J. Liu, and C. Wen Chen. 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*, pages 4535–4544, 2017.
- [72] L. Mai, H. Jin, and F. Liu. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 497–506, 2016.
- [73] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *2011 international conference on computer vision*, pages 1784–1791. IEEE, 2011.
- [74] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.
- [75] H. Müller, P. Clough, T. Deselaers, B. Caputo, and I. CLEF. Experimental evaluation in visual information retrieval. *The Information Retrieval Series*, 32:1–554, 2010.
- [76] N. Murray and A. Gordo. A deep architecture for unified aesthetic prediction. *arXiv preprint arXiv:1708.04890*, 2017.
- [77] N. Murray, L. Marchesotti, and F. Perronnin. AVA: a large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, Providence, Rhode Island, 2012. IEEE.
- [78] H.-V. Nguyen and J. Vreeken. Non-parametric jensen-shannon divergence. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 173–189. Springer, 2015.
- [79] D. W. Otter, J. R. Medina, and J. K. Kalita. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

- [80] K. Park, S. Hong, M. Baek, and B. Han. Personalized image aesthetic quality assessment by joint regression and ranking. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1206–1214. IEEE, 2017.
- [81] N. M. Razali, Y. B. Wah, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1): 21–33, 2011.
- [82] R. Reber, N. Schwarz, and P. Winkielman. Processing fluency and aesthetic pleasure: Is beauty in the perceiver’s processing experience? *Personality and social psychology review*, 8(4):364–382, 2004.
- [83] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
- [84] J. Ren, X. Shen, Z. Lin, R. Mech, and D. J. Foran. Personalized image aesthetics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 638–647, 2017.
- [85] F. Ribeiro, D. Florencio, and V. Nascimento. Crowdsourcing subjective image quality evaluation. In *2011 18th IEEE International Conference on Image Processing*, pages 3097–3100. IEEE, 2011.
- [86] H. Roy, T. Yamasaki, and T. Hashimoto. Predicting image aesthetics using objects in the scene. In *Proceedings of the 2018 International Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia*, pages 14–19, 2018.
- [87] R. Schifanella, M. Redi, and L. M. Aiello. An image is worth more than a thousand favorites.
- [88] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61: 85–117, 2015.
- [89] K. Schwarz, P. Wieschollek, and H. P. Lensch. Will people like your image? learning the aesthetic space. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2048–2057, Lake Tahoe, United States, 2018. IEEE.
- [90] S. Shekhar, D. Singal, H. Singh, M. Kedia, and A. Shetty. Show and recall: Learning what makes videos memorable. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2730–2739, 2017.

- [91] K. Sheng, W. Dong, H. Huang, C. Ma, and B.-G. Hu. Gourmet photography dataset for aesthetic assessment of food images. In *SIGGRAPH Asia 2018 Technical Briefs*, pages 1–4. 2018.
- [92] K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang, and B.-G. Hu. Attention-based multi-patch aggregation for image aesthetic assessment. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 879–886, 2018.
- [93] K. Sheng, W. Dong, M. Chai, G. Wang, P. Zhou, F. Huang, B.-G. Hu, R. Ji, and C. Ma. Revisiting image aesthetic assessment via self-supervised feature learning. In *AAAI*, pages 5709–5716, 2020.
- [94] R. N. Shepard. Recognition memory for words, sentences, and pictures. *Journal of verbal Learning and verbal Behavior*, 6(1):156–163, 1967.
- [95] E. Siahaan, A. Hanjalic, and J. Redi. A reliable methodology to collect ground truth data of image aesthetic appeal. *IEEE Transactions on Multimedia*, 18(7):1338–1350, 2016.
- [96] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [97] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [98] M. Suchecki and T. Trzciski. Understanding aesthetics in photography using deep convolutional neural networks. In *2017 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 149–153. IEEE, 2017.
- [99] W.-T. Sun, T.-H. Chao, Y.-H. Kuo, and W. H. Hsu. Photo filter recommendation by category-aware aesthetic learning. *IEEE Transactions on Multimedia*, 19(8):1870–1880, 2017.
- [100] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [101] H. Talebi and P. Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.
- [102] X. Tang, W. Luo, and X. Wang. Content-based photo quality assessment. *IEEE Transactions on Multimedia*, 15(8):1930–1943, 2013.

- [103] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 1(8), 2015.
- [104] A. Tifentale and L. Manovich. Competitive photography and the presentation of the self. In *Exploring the Selfie*, pages 167–187. Springer, 2018.
- [105] H. Tong, M. Li, H.-J. Zhang, J. He, and C. Zhang. Classification of digital photos taken by photographers or home users. In *Pacific-Rim Conference on Multimedia*, pages 198–205. Springer, 2004.
- [106] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3d: generic features for video analysis. *CoRR*, *abs/1412.0767*, 2(7):8, 2014.
- [107] C. Wan and X. Tian. A small scale multi-column network for aesthetic classification based on multiple attributes. In *International Conference on Neural Information Processing*, pages 922–932, Guangzhou, China, 2017. Springer.
- [108] G. Wang, J. Yan, and Z. Qin. Collaborative and attentive learning for personalized image aesthetic assessment. In *IJCAI*, pages 957–963, 2018.
- [109] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.
- [110] W. Wang and R. Deng. Modeling human perception for image aesthetic assessment. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1029–1033. IEEE, 2019.
- [111] W. Wang and J. Shen. Deep cropping via attention box prediction and aesthetics assessment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2186–2194, 2017.
- [112] W. Wang, M. Zhao, L. Wang, J. Huang, C. Cai, and X. Xu. A multi-scene deep learning model for image aesthetic evaluation. *Signal Processing: Image Communication*, 47: 511–518, 2016.
- [113] W. Wang, J. Su, L. Li, X. Xu, and J. Luo. Meta-learning perspective for personalized image aesthetics assessment. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1875–1879. IEEE, 2019.

- [114] W. Wang, S. Yang, W. Zhang, and J. Zhang. Neural aesthetic image reviewer. *IET Computer Vision*, 13(8):749–758, 2019.
- [115] Z. Wang, S. Chang, F. Dolcos, D. Beck, D. Liu, and T. S. Huang. Brain-inspired deep networks for image aesthetics assessment. *arXiv preprint arXiv:1601.04155*, 2016.
- [116] Z. Wang, D. Liu, S. Chang, F. Dolcos, D. Beck, and T. Huang. Image aesthetics assessment using deep chatterjee’s machine. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 941–948. IEEE, 2017.
- [117] L.-K. Wong and K.-L. Low. Saliency-enhanced image aesthetics class prediction. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 997–1000. IEEE, 2009.
- [118] J. Ye, J. Li, M. G. Newman, R. B. Adams, and J. Z. Wang. Probabilistic multigraph modeling for improving the quality of crowdsourced affective data. *IEEE transactions on affective computing*, 10(1):115–128, 2017.
- [119] J. Yu, C. Cui, L. Geng, Y. Ma, and Y. Yin. Towards unified aesthetics and emotion prediction in images. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2526–2530. IEEE, 2019.
- [120] W. Zhang, G. Zhai, X. Yang, and J. Yan. Hierarchical features fusion for image aesthetics assessment. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3771–3775. IEEE, 2019.
- [121] X. Zhang, X. Gao, W. Lu, and L. He. A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction. *IEEE Transactions on Multimedia*, 21(11):2815–2826, 2019.
- [122] X. Zhang, X. Gao, W. Lu, Y. Yu, and L. He. Fusion global and local deep representations with neural attention for aesthetic quality assessment. *Signal Processing: Image Communication*, 78:42–50, 2019.
- [123] Y. Zhou, X. Lu, J. Zhang, and J. Z. Wang. Joint image and text representation for aesthetics analysis. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 262–266, Amsterdam, Netherlands, 2016. ACM.
- [124] G. Y. Zou. Toward using confidence intervals to compare correlations. *Psychological methods*, 12(4):399, 2007.

Titre: Évaluation de la qualité esthétique d'image par apprentissage profond

Mots clés: Esthétique, Traitement des Images / Vidéos, Apprentissage Automatique, Réseau Neuronal, Évaluation de la Qualité Visuelle

Résumé: Avec le développement des dispositifs de capture et d'Internet, les utilisateurs ont accès à une quantité croissante d'images. L'évaluation de l'esthétique visuelle a des applications importantes dans plusieurs domaines, de la recherche et de la recommandation d'images à leur amélioration. L'évaluation de la qualité esthétique des images vise à déterminer la beauté d'une image pour les observateurs humains. De nombreux problèmes dans ce domaine ne sont pas bien étudiés, notamment la subjectivité de l'évaluation de la qualité esthétique, l'explication de l'esthétique et la collecte de données annotées par des humains. Les méthodes conventionnelles pour la prédiction de la qualité esthétique des images visent à prédire le score moyen ou la classe esthétique d'une image. Cependant, la prédiction esthétique est intrinsèquement subjective, et des images ayant des notes moyennes ou des classes esthétiques similaires peuvent présenter des niveaux de consensus très différents selon les évaluateurs humains. Des travaux récents ont traité de la subjectivité esthétique en prédisant la distribution des scores humains, mais la prédiction de la distribution n'est pas directement interprétable en termes de subjectivité, et pourrait être sous-optimale par rapport à l'estimation directe des descripteurs de subjectivité calculés à partir de scores de vérité terrain. En outre, les étiquettes des bases de données existants sont souvent bruitées, incomplètes ou ne permettent pas d'effectuer des tâches plus sophistiquées, comme comprendre pourquoi une image est belle ou non pour un observateur humain.

Dans cette thèse, nous proposons d'abord plusieurs mesures de la subjectivité, allant de simples mesures statistiques telles que l'écart-type des scores, à des nouveaux descripteurs inspirés de la théorie de l'information. Nous évaluons la performance de prédiction de ces

mesures lorsqu'elles sont calculées à partir de distributions de scores prédites et lorsqu'elles sont directement apprises à partir de données de vérité terrain. Nous constatons que cette dernière stratégie donne en général de meilleurs résultats. Nous utilisons également la subjectivité pour améliorer la prédiction des scores esthétiques, en montrant que les mesures de subjectivité inspirées par la théorie de l'information donnent de meilleurs résultats que les mesures de subjectivité inspirées par la statistique.

Ensuite, nous proposons une base de données appelée « Esthétique Visuelle Explicable » (EVA), qui contient 4070 images avec au moins 30 votes par image. Les votes dans EVA ont été obtenus en ligne en utilisant une approche plus disciplinée inspirée des meilleures pratiques d'évaluation de la qualité. La base offre également des caractéristiques supplémentaires, telles que le degré de difficulté de l'évaluation de la note esthétique, la notation de 4 attributs esthétiques complémentaires, ainsi que l'importance relative de chaque attribut dans la formation d'une opinion esthétique. La base de données a été mise à disposition du public et devrait contribuer aux futures recherches sur la compréhension et la prévision de la qualité esthétique visuelle.

En outre, nous avons étudié l'explicabilité de l'évaluation de la qualité esthétique des images. Une analyse statistique sur EVA démontre que les attributs collectés et leur importance relative peuvent être combinés de manière linéaire pour expliquer efficacement les scores moyens globaux des opinions esthétiques. Nous avons constaté que la subjectivité a une corrélation limitée avec la difficulté personnelle moyenne de l'évaluation esthétique, et que la provenance géographique, le niveau photographique et l'âge du sujet affectent de manière significative l'évaluation esthétique de l'utilisateur.

Title: Image Aesthetic Quality Assessment Based on Deep Neural Networks

Keywords: Aesthetics, Image/Video Processing, Machine Learning, Neural Network, Visual Quality Assessment

Abstract: With the development of capture devices and the Internet, people access to an increasing amount of images. Assessing visual aesthetics has important applications in several domains, from image retrieval and recommendation to enhancement.

Image aesthetic quality assessment aims at determining how beautiful an image looks to human observers. Many problems in this field are not studied well, including the subjectivity of aesthetic quality assessment, explanation of aesthetics and the human-annotated data collection. Conventional image aesthetic quality prediction aims at predicting the average score or aesthetic class of a picture. However, the aesthetic prediction is intrinsically subjective, and images with similar mean aesthetic scores/class might display very different levels of consensus by human raters. Recent work has dealt with aesthetic subjectivity by predicting the distribution of human scores, but predicting the distribution is not directly interpretable in terms of subjectivity, and might be sub-optimal compared to directly estimating subjectivity descriptors computed from ground-truth scores. Furthermore, labels in existing datasets are often noisy, incomplete or they do not allow more sophisticated tasks such as understanding why an image looks beautiful or not to a human observer.

In this thesis, we first propose several measures of subjectivity, ranging from simple statistical measures such as the standard deviation of the scores, to newly proposed descriptors in-

spired by information theory. We evaluate the prediction performance of these measures when they are computed from predicted score distributions and when they are directly learned from ground-truth data. We find that the latter strategy provides in general better results. We also use the subjectivity to improve predicting aesthetic scores, showing that information theory inspired subjectivity measures perform better than statistical inspired subjectivity measures.

Then, we propose an Explainable Visual Aesthetics (EVA) dataset, which contains 4070 images with at least 30 votes per image. EVA has been crowd-sourced using a more disciplined approach inspired by quality assessment best practices. It also offers additional features, such as the degree of difficulty in assessing the aesthetic score, rating for 4 complementary aesthetic attributes, as well as the relative importance of each attribute to form aesthetic opinions. The publicly available dataset is expected to contribute to future research on understanding and predicting visual quality aesthetics.

Additionally, we studied the explainability of image aesthetic quality assessment. A statistical analysis on EVA demonstrates that the collected attributes and relative importance can be linearly combined to explain effectively the overall aesthetic mean opinion scores. We found subjectivity has a limited correlation to average personal difficulty in aesthetic assessment, and the subject's region, photographic level and age affect the user's aesthetic assessment significantly.