PH.D SCHOOL
UNIVERSITY OF MILANO-BICOCCA

Department of Informatics, Systems and Communication
PhD Program in Computer Science - XXXIV Cycle

Image Collection Management using Convolutional Neural Networks

Ph.D. Dissertation of: Marco Leonardi

Supervisor: Prof. Raimondo Schettini
Co-Supervisor: Prof. Paolo Napoletano
Co-Supervisor: Dr. Alessandro Rozza
Tutor: Prof. Fabio Stella
Ph.D. Coordinator: Prof. Leonardo Mariani

Academic Year 2020-2021

Acknowledgements

The journey of doing a PhD has been a life-changing experience. It would not have been possible to do without the support and guidance that I received from countless people.

First and foremost I am extremely grateful to my supervisor, Raimondo Schettini for his invaluable advice, and for believing in me during my PhD study.

I would like to acknowledge Paolo Napoletano, my co-supervisor, for the precious help, insightful feedback, and patience he showed during these years.

Further, I would like to thank Alessandro Rozza, my co-supervisor, for his advice and suggestions.

I would like to thank Luigi Celona, for his friendship with me and for his precious suggestions.

In addition, I would like to acknowledge my colleagues Simone Zini, Davide Marelli, and Alessio Albé for the friendly and stimulating environment that have been established in the Imaging and Vision Laboratory. I would like to thank the acquired colleagues and all the people that have been in Imaging and Vision Laboratory during these years with which I shared many joyful moments.

I would like to express my gratitude to my family, for their presence and support, without which it would not have been possible. I owe it to you.

Finally, I would like to acknowledge all my friends, who made me smile in the bad moments, and Sara, my love.

Abstract

Almost everyone carries a high-quality camera in their smartphone and uses it to communicate with other individuals and for the last two decades, people are increasingly making use of images and videos in their transportable communication. As the prices of the storage are decreasing, the number of photos stored is increasing, leading to collections of images whose sizes begin to be a barrier for relieving the captured moments and exploring them. We are submerged by images.

In order to ease the problem of oversized image collections, methods that aim to select a subset of photos that best represents them have been designed and proposed in the literature. Those methods typically rely upon the prediction of perceptual features such as, for example, the image quality, aesthetics, and memorability, to select the best images.

This thesis starts from the fundamental image properties that guide the image selection, respectively the image quality and image aesthetics. First, the perceived image quality assessment is investigated in an anomaly detection manner, contrary to the most common regression task. This is because rather than predict a score that best correlates to the average human opinion, being able to distinguish good quality images from bad ones, is more suitable for the image collection management problem, furthermore, it requires fewer images to tune the model. Then the problem of automatic assessment of image aesthetics is introduced. In the beginning, presenting a method that learns the aesthetics of a picture on the basis of the prediction of aesthetics-related attributes. Then, a new solution that takes into account the semantic content, the artistic style, and the composition of the image is presented.

One of the reasons people take photos is to capture important situations to recall them later on, usually with the intention of afterwards sharing their photos with other people like friends or family members. Photos can be seen as a concrete link between our memories and experienced events. Image memorability can be helpful in the organization of the selected images to better bind the memory of experienced events and the taken images. To this end in this thesis, a method for the estimation of still image memorability is presented. In particular, the proposed method goes in the direction of breaking down the intrinsic image properties that influence the memorability of the pictures.

Image collections tend to have several similar images. This is because to ensure the best shot, people usually take a series of photos of the same scene. To guarantee a diverse and representative selection of images from a large collection, this thesis concludes by proposing a flexible and innovative framework that can be used to both explore large-

scale image datasets and to summarize photo albums. The proposed method is designed to exploit different aspects of the images, such as the scene category, image quality, and image aesthetics.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Why image collection management

Images are the oldest form of transportable communication (i.e. communication that can be "transported" across time and space). The most antique image ever found is a life-sized picture of a wild pig. It dates at least 45,500 years and origins from Sulawesi, an island in Indonesia. This kind of prehistoric artwork can still nowadays communicate to us hints like the deep symbolic significance of Sulawesi warty pigs in the ancient hunting culture. The first photo was ever taken, a view from the window at Le Gras, dates back to 1825 by French inventor Joseph Nicéphore Niépce. It was the kick-starter of photography.
Nowadays, in the era of digital communication, images are increasingly the means by which people communicate. According to the annual Internet Trends Report of 2019 from Bond[1], on Twitter[2], a platform that should be mostly text-only, today more than half of the posts involves images, video or other media.

Nowadays taking a photo has become such easy as normal in nearly every situation of our life. We take photos to capture moments, from the most personal events to the less trivial situations of our daily routine. To remember things or even document situations. Taking a photo has become part of our everyday life: images are a way for documenting our lifetime, instead of relying on the individual's perspective, a photograph can frame a moment in time.

Almost everyone carries a high-quality camera in their smartphone and uses it to communicate with other individuals and for the last two decades, people are increasingly making use of images and videos in their transportable communication. Encouraged by social media platforms to promote ourselves through digital content, and the declining storage device prices as well have lead to a tremendous increase in personal digital content. It is estimated that trillions of photos are captured globally each year [3].

In the modern era of smartphone photography, the ambition of taking a good image

[1] www.bondcap.com/report/itr19
[2] www.twitter.com
[3] www.riseaboveresearch.com

has transfer the effort made by the people, from the moment of taking the picture, to the stage of selecting the best shoot. Indeed it's becoming a common practice to shoot the same scene several times, deferring the choice of the best image to later, rather than spend time to ensure the best possible single shot during the process of capturing the moment.

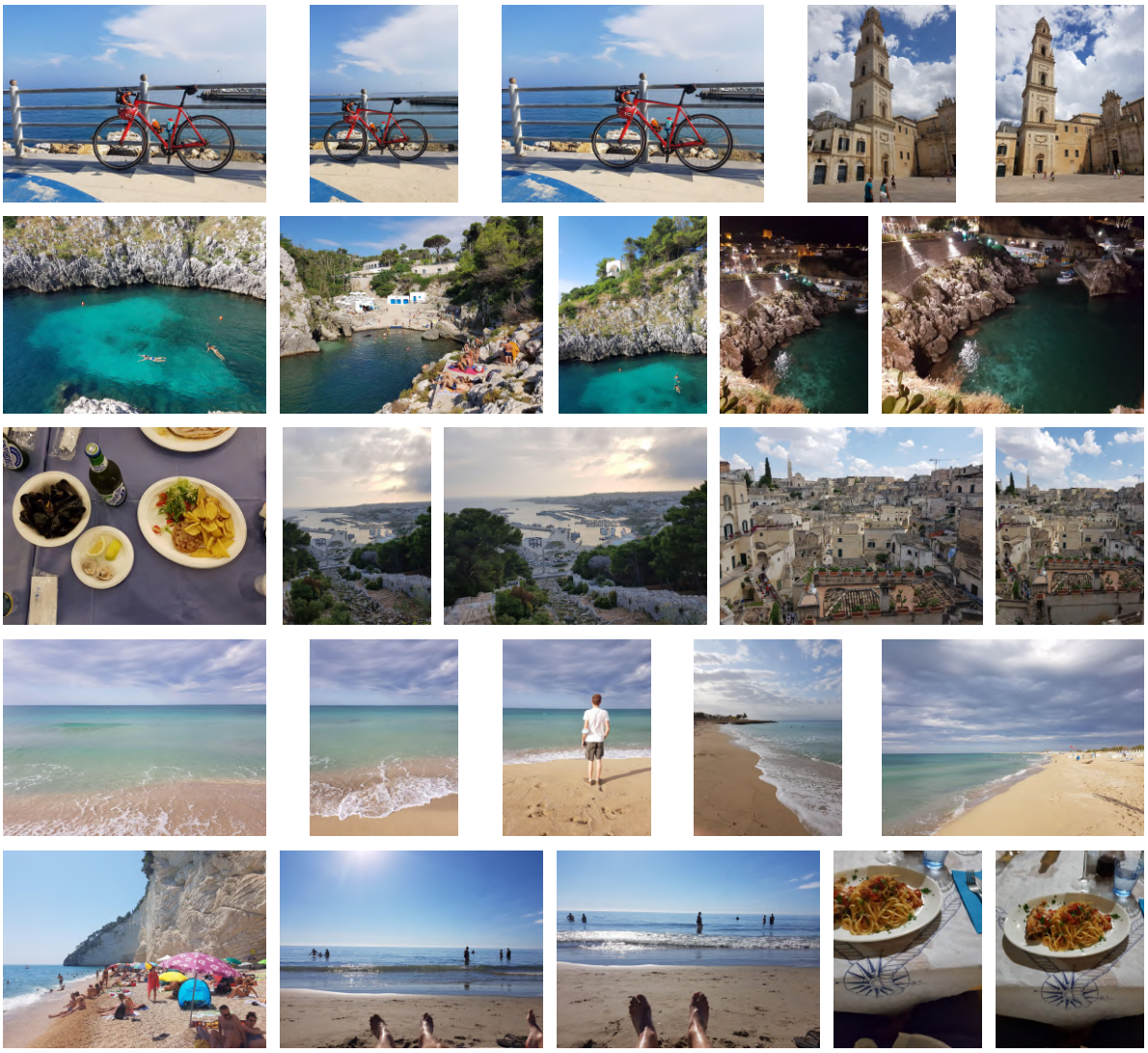


Figure 1.1: Example of collection of captured photos with several repeated elements.

One of the major issues, that emerged as a consequence of accumulating large collections of captured photos, is that reliving the moments imprinted in the pictures, or simply navigate through them, is becoming a tedious and unappealing task. This contributes to a form of digital forgetting: users are spurred by the number of photos, and consequently, those photos become rarely accessed and enjoyed again in the future. In figure 1.1 is

reported as an example, a subset of a collection of images with several repeated elements. The aim of automatic image collection management is to select the best images that represent the whole group highlighting the various moments without losing information.

In response to these challenges, which emerged with large images collections, automatic image collection management is becoming of paramount importance: the aim is to be able to automatically manage huge amount of images preserving the best and most valuable pictures.

Over the last years, various methods for automatic photo selection have been proposed. Recent methods, as a result of the popularity of machine learning, moved from the employ of visual features of images (i.e. colour histogram) to property that represents subjective characteristics, attempting to apprehend how users perceive a picture (i.e. image quality). Although progress has been made in this field, the problem of image collection management remains open.

1.2 The role of perceptual properties

The process of selecting a subset of photos from a larger collection, which best represent the information in the whole, is guided by perceptual features.

The act of selecting, from a phenomenological point of view, involves an interest in some pictures rather than others. This interest is related to various perceptual properties, such as *image aesthetics* [69], *image quality* [70] and *image memorability* [68].

Humans make judgements and choices every day based on their inner aesthetic responses to aspects of the surrounding world. We make a decision, e.g., we buy this poster rather than that one because we like its graphic composition, or we simply choose to sit facing this direction rather than that one in the park because we find the view more pleasurable [99]. One of the key determining factors that guide humans judgements in image selection is *image aesthetic*. The aesthetics of a picture can be defined as the measure of appreciation of the beauty of an image. Image aesthetics is a subjective property that depends on the viewer's preferences, experiences, and skills as a photographer. Despite this, the occurrence of specific factors or patterns objectively makes an image more appealing than others. Image aesthetics can be influenced by several factors [35, 109, 53]. When people compare similar images tends to prefers brighter colors, sharp pictures and

(a) (b) (c) (d)

Figure 1.2: Sample pairs of similar images with different aesthetics and quality scores. Respectively high (a) image aesthetics and low (b). High image quality (c),low (d).

clearer and closer photos showing more detail. On the other hand, people usually dislike images with dark lighting or washed colors [18]. Figure 1.2a and figure 1.2b shows the same subject with different composition styles and colors. Figure 1.2a is preferable to 1.2b in terms of aesthetics since figure 1.2a has brighter colors and a better image style composition, while in figure 1.2b colors are washed and the lights are darker.

Our conscious awareness is strongly correlated with the vision: more than 30% of cortical neurons are devoted to vision [24]. As visual creatures, we are drawn to visual realism and naturalness [10]. Therefore the demand for high-quality images is constantly increasing. Despite cameras and sensors manufacturers are trying to satisfy this demand with more sophisticated and precise technologies, it is not rare to shoot pictures that have a low perceived visual quality. Poor external conditions, such as low light environment, backlight scenes, or moving objects, alongside erroneous capturing settings, like exposure, ISO (camera setting that brightens or darken a photo), and aperture, could cause annoying image artefacts and distortions that lead to an unsatisfactory perceived visual quality. *Image quality* plays an important role in the selection of images, it is widely acknowledged that people prefer pristine images rather than pictures affected by distortions. Figure 1.2c and figure 1.2d shows the same subject with different perceived image quality. Immediately catches the eye that the erroneous capturing settings affected the figure 1.2d with motion blur and burned area. Instead figure 1.2c appears to be flawless with respect to the perceived image quality. Therefore 1.2c is preferable than 1.2d in terms of image quality.

(a) (b) (c) (d)

Figure 1.3: Examples of high memorable images (a-b) and low memorable images (c-d) from the LaMem dataset [58].

One of the reasons people take photos is to capture important situations to recall them later on, usually with the intention of afterwards share their photos with other people like friends or family members [85]. Image collections are a sort of diary of our past experiences, like holidays, special events and so forth. From a psychological point of view, taking photos improves memory for visual aspects of an experience. Taking pictures mentally similarly heightens visual memory [3]. Photos can be seen as a concrete link between our memories and experienced events. Moreover, humans have the remarkable ability to remember whether they have seen an image before, even after seeing thousands of images, each only once and only for a few seconds [117]. This peculiar property of some pictures to stick in our minds better than others is called *image memorability* [58].
Image memorability can be helpful in the organization of the selected images to better

bind the memory of experienced events and the taken images. In the figure 1.3(a-b) are represented some samples of high memorable images, while in the figure 1.3(c-d) are depicted examples of low memorable images.

1.3 Diversity

There are several reasons why there are plenty of similar images in our collections. Sometimes the dynamicity of the scene forces us to take a photo burst rather than a single picture. Other times, to ensure the best photo, rather than spend time to evaluate the composition of the image or other factors that can influence the aesthetics of the final result, people prefer to take several images of the same scene, postponing the selection of the best shot to later. For instance, during a sportive activity, a typical user would shoot hundreds of images, many of which are repetitive and redundant. In the figure 1.4 is reported a typical series of pictures of the same scene.

(a) (b) (c) (d) (e)

Figure 1.4: Series of shots of a particular scene.

With the image quality, we can get rid of those annoying images affected by artefacts and consider only the pristine ones. By the image aesthetics instead, we can select the most beautiful pictures from our collection. While with the image memorability we can reorder them so that the most memorable pictures come first. Unfortunately, those features do not ensure that the selection is devoid of similar pictures.

Since the differences are usually small (e.g. change of perspective, movement of the subject), similar images may also have similar properties, therefore they may all be eligible. Consequently, the resulting set that represents the given collection of images may suffer from low entropy. To this end, is fundamental the concept of image diversity. During the process of summarizing a set of pictures, the resulting subset of images should capture every moment, without annoying repeated elements.

1.4 Thesis overview

Exploring large collections of images is nowadays a problem more than ever: images are everywhere and the stored photos are increasing day by day. The main focus of this thesis is to tackle the problem of image collection management using convolutional neural networks. To this end, this thesis follows a top-down approach to the problem, focusing on both the general task and on the single problems that are implied.

When surfing a wide collection of images, systems that can automatically select the most relevant pictures are of paramount importance to ease the charge. Sometimes the very massiveness of the collection is itself a deterrent to the users who want to explore them. An image collection management system can be seen as a system that takes as input a collection of images and returns as output the most representative images of the given set. In the figure 1.5 is reported a conceptualization of an image collection management system.

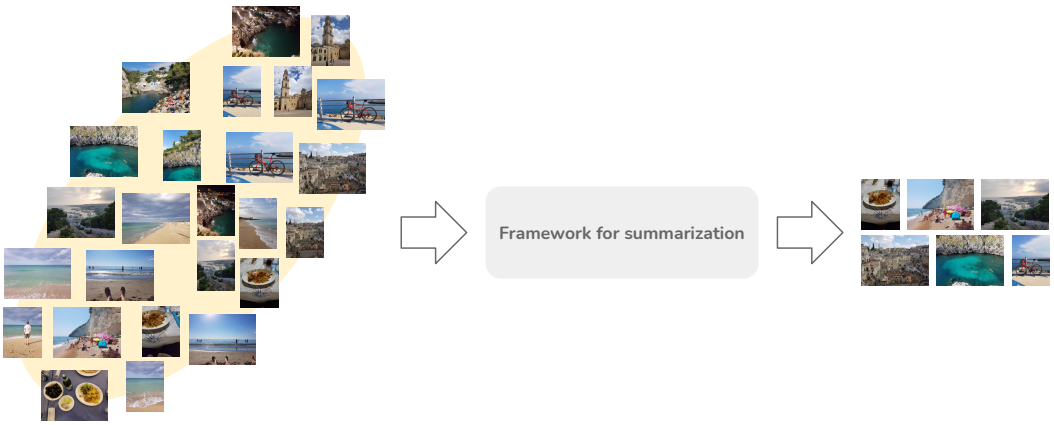


Figure 1.5: Schematic overview of an image collection management system.

The process of summarizing a collection of images can be broken down into single subproblems. For example, when we have to create a photo album of our vacation, a possible strategy is to divide the images into groups and then, for each of the groups, select the most representative image. Similarly, this thesis tackles the problem of image collection management in a modular way. Therefore the first task should be to group images into homogeneous groups as depicted in the figure 1.6. The criterion by which the images are grouped is related to the kind of image collection the system is applied to. Some examples of criteria to group images are the image content, the timestamp, or the similarity between two images.

The subsequent task is the selection of the most representative image from each of the previously proposed groups. As stated before, the act of selecting involves an interest in some pictures rather than others. Interest that can be modelled with various perceptual properties, such as *image aesthetics*, *image quality* and *image memorability*. Therefore each of these properties has to be considered. An example of the process of selecting the

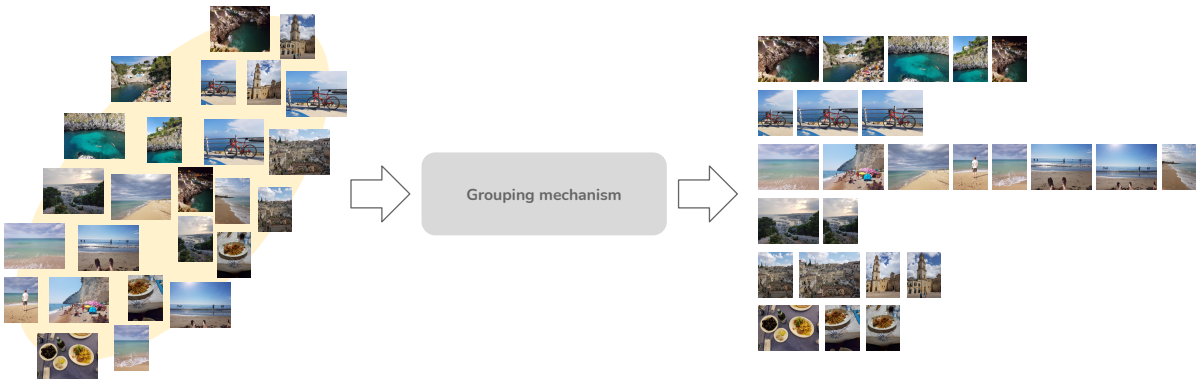


Figure 1.6: Given a collection of images, an image collection management system should be able to group similar images.

best images from each group is reported in the figure 1.7, while in figure 1.8 is highlighted an example the procedure for selecting the most interesting picture according to the *image aesthetics* and *quality*.

Figure 1.7: For each of the group of images, the most representative image is selected.

Given the modularity of the approach to the task of image collection management, this thesis considers each of the submodules as stand-alone problems with respect to the state of the arts, therefore each of the submodules has its own literature.

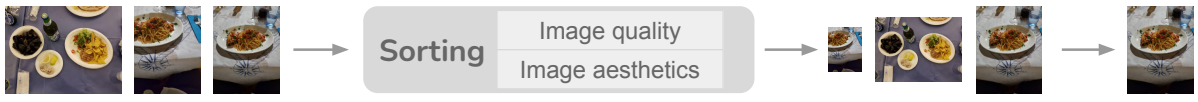


Figure 1.8: For a given group of images, pictures are ordered with respect to the *image aesthetics* and the *image quality*. The most representative one is then selected.

This thesis starts with the submodules that guide the image selection, therefore the

image quality and *image aesthetics*, comparing the produced work with existing solutions in the state of the arts. Later the attention is shifted to the module which leads the ordering of the selected images, the *image memorably*. Finally, the module in charge of maximising the diversity is introduced presenting a solution to the task of image collection management joining all the previous modules.

1.5 Significant contributions

Given the modularity of this thesis, the contributions are distributed over the aforementioned submodules, in particular with respect to the *image quality*, *image aesthetics*, *image memorability* and *image diversity*. The state of the arts is therefore analyzed for each of these problems.

The thesis begins with the perceived image quality assessment. After presenting the review of the state of the arts, a solution that aims to quantitatively represent the human perception of quality is introduced. Although the problem of quality assessment is typically addressed as a regression task, in this thesis the perceived image quality assessment is reformulated in an anomaly detection manner. This is because rather than predict a score that best correlates to the mean opinion score, being able to distinguish good quality images from bad ones, is more suitable for the image collection management problem.

The focus is later shifted to automatic assessment of image aesthetics. In the beginning, a preliminary study is made by learning the aesthetics score on the basis of the prediction of aesthetics-related attributes. Then, over the assumption that aesthetics is inherently a subjective property influenced by certain factors such as the semantic content of the image, the attributes describing the artistic aspect, the photographic setup used for the shot, etc. A new method for the automatic prediction of the aesthetics of an image that is based on the analysis of the semantic content, the artistic style and the composition of the image is presented. For each of the two studies, the state of the art is revised considering the problem that is related to them.

Still image memorability estimation is then introduced, a relatively new and recent problem as the review of the state of the arts points out. The proposed method goes in the direction of breaking down the intrinsic image properties that influence the memorability of the pictures. These intrinsic properties are respectively *what* kind of objects and scenes are present and *what* are their characteristics, but also by extrinsic factors such as the image locations *where* humans focus their attention.

Finally, the concept of image diversity is explored. After reviewing the state of the art, a flexible and innovative framework that can be used both to explore large scale image datasets and to summarize photo albums is proposed. The method first separates images with respect to the scene category. Then, for each category, aims to select the most diverse and beautiful images that represents that category. The presented pipeline is based on features extracted from a Convolutional Neural Network to assess the diversity and perceptual properties like the *image quality* and *image aesthetics* to ensure the selection of pristine images.

Chapter 2

No Reference, Opinion Unaware Image Quality Assessment by Anomaly Detection

2.1 Introduction

Even if cameras and sensors are becoming increasingly sophisticated and precise, it is not rare to shoot pictures that have a low perceived visual quality. Poor external conditions, such as low light environment, backlight scenes, or moving objects, alongside erroneous capturing settings, like exposure, ISO (camera's sensitivity to light), and aperture, could cause annoying image artifacts and distortions that lead to an unsatisfactory perceived visual quality. Being able to automatically distinguish good quality images from the bad ones can help various types of applications to prune the input set of images, like automatic photo album creation, or even help users to discard bad quality images from their personal collection to save space and time for revisiting those pictures.

Even if subjective assessment is the most accurate criterion for image quality, it is not possible to perform it over big collections in a relatively small amount of time. Without taking into account the expensiveness and the cumbersomeness of manually evaluate pictures, automatic Image Quality Assessment (IQA) algorithms have been widely studied in recent years.

The perceptual quality of an image is a subjective measure and it is usually defined as the mean of the individual ratings of perceived quality assigned by human subjects (Mean Opinion Score—MOS). Given an image, IQA systems are designed to automatically estimate the quality score. Existing IQA methods can be classified into three major categories: full-reference image quality assessment (FR-IQA) algorithms, e.g., [1, 9, 32, 45, 101, 128, 2], reduced-reference image quality assessment (RR-IQA) algorithms, e.g., [129, 107, 80, 73], and no-reference/blind image quality assessment (NR-IQA) algorithms, e.g., [86, 79, 76, 139, 124]. FR-IQA methods compare the distorted image with respect to the reference image in order to predict a quality score, and because of that, they

require both the original image and the corrupted one. RR-IQA algorithms assess the image quality by exploiting partial information of the corresponding reference image. The quality prediction is the result of a comparison between features extracted from the reference and the image under test. NR-IQA algorithms estimate the image quality solely on the distorted picture without the need of the reference image.

In this thesis, the focus is specifically on the no-reference opinion-unaware image quality assessment. Driven by the need of being able to distinguish good quality images from the bad ones, rather than predict a score that best correlate to the mean opinion score, the problem is tackled by an anomaly detection approach.

The main contributions of this thesis with respect to the problem of image quality assessment are:

- It is proposed an anomaly detection no-reference opinion-unaware method capable of estimating the image quality by exploiting the correlations between feature maps extracted by means of a convolutional neural network.
- The effectiveness of the method is demonstrated with respect to one no-reference opinion-unaware image quality assessment and two opinion-aware methods, on three databases containing real distorted images.
- It is proposed different way to evaluate NR-IQA methods based on their capabilities to discriminate good quality images from the bad ones.
- The advantages of combining the correlations between feature maps with an anomaly detection method are shown.

2.2 Related work

The majority of the existing methods on NR-IQA are designed to be trained on datasets composed of mean opinion scores in addition to the distorted images. Ye et al. in [133] proposed an unsupervised feature learning framework named Codebook Representation for No-Reference Image Assessment (CORNIA). They rely on the use of codebook, a collection of different type of features which encode the quality of an image. The framework can be summed by four major steps: local feature extraction, codebook construction, local feature encoding and feature pooling. In [94], Moorthy et al. present the Distortion Identification-based Image Verity and INtegrity Evaluation index (DIIVINE), a two-stage framework for estimating quality, based on the hypothesis that statistical properties of images change in presence of distortions. These two stages are respectively the distortion identification followed by distortion-specific quality assessment. The core of the method uses a Gaussian scale mixture to model neighboring wavelet coefficients to then extract the statistical description of the distortion in the input image. Lately, most of the works in the state of the arts focus on the use of deep learning: Bianco et al. in [7] propose DeepBIQ, a CNN pretrained on the union of ImageNet [108] and Places [142] datasets

fine-tuned for the image quality task. In [15] Celona et al. schematically illustrate building blocks that are implemented and combined in different ways on CNN-based framework for evaluating image quality of consumer photographs.

Nevertheless there are methods that further relax the constraint posed by the previously introduced NR-IQA algorithms, removing the requirement of having human subjective scores, thus leading in two new sub category of the no-reference image quality assessment algorithms namely opinion-aware and opinion-unaware/completely-blind.

Since they do not require training samples of distortions nor of human subjective scores, opinion-unaware methods are usually robust generalization capability. Zhang et al. in [136] propose ILNIQE, an opinion-unaware no-reference image quality assessment algorithm based on natural scene statistics (NSS). In particular, they learn a multivariate Gaussian model of image patches from a collection of pristine natural images. They then compute the overall quality score average pooling the quality of each image patch of a given image using a Bhattacharyya-like distance over the learned multivariate Gaussian model.

It is well known that convolutional neural networks have the capability of intrinsically encode quality of the images in their deep visual representation [138, 8]. In [38], Gatys et al. represent textures by the correlations between feature maps in several layers of a convolutional neural network, and show that across layers the texture representations increasingly capture the statistical properties of natural images. Then in [39] they exploit this correlations between feature maps to create a loss function that measures the difference in style between two images. Other studies reveals how this kind of correlations can be used to classify style of the images [41, 22]. In this thesis, is used for the first time the correlations between feature maps to the task of no-reference opinion-unaware image quality assessment.

2.3 Proposed method

The aim of the proposed method is to estimate the quality score of a given picture exploiting the correlations between the feature maps coming from a VGG16 [114] trained on the Imagenet dataset [108]. In particular, the arithmetic mean of this correlations, and a score resulting from an anomaly detection method, on the intra-layer correlation represented by the Gram matrix, are combined. The final image quality score is given by the sum of these two metrics after applying min-max scaling to them. A brief overview of the method is depicted in Figure 2.1.

2.3.1 Intra-Layer Correlation

The intra-layer correlation has been firstly introduced by Gatys et al. [38] as representative of the texture, and since then it has been used in several other studies: as a loss [39, 54] for style transfer, or to classify the style of the images [41, 22]. This property can be defined as the correlation between the feature maps in a given layer and it can be done

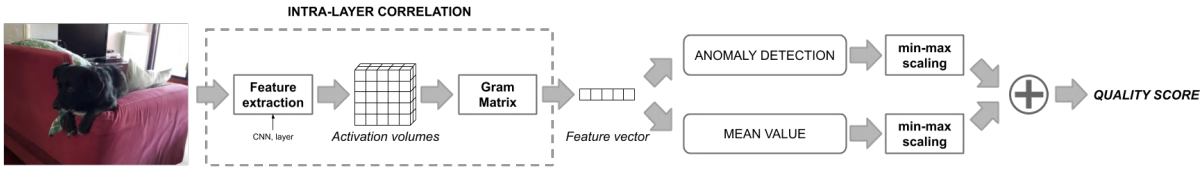


Figure 2.1: Schematic view of the proposed method. The intra-layer correlation is computed by the Gram matrix over the activation volumes of a Convolutional Neural Network (CNN). Then the *abnormality* and the average of the correlation are computed before applying the min-max scaling on both of them. In the end, the two metrics are summed resulting in the predicted image quality score.

by calculating the Gram matrix of the feature maps, which is defined as all possible inner products of the feature maps. The Gram matrix has also been used in several other domains and applications: from computing linear independence of a set of vectors to represents kernel functions as in [65].

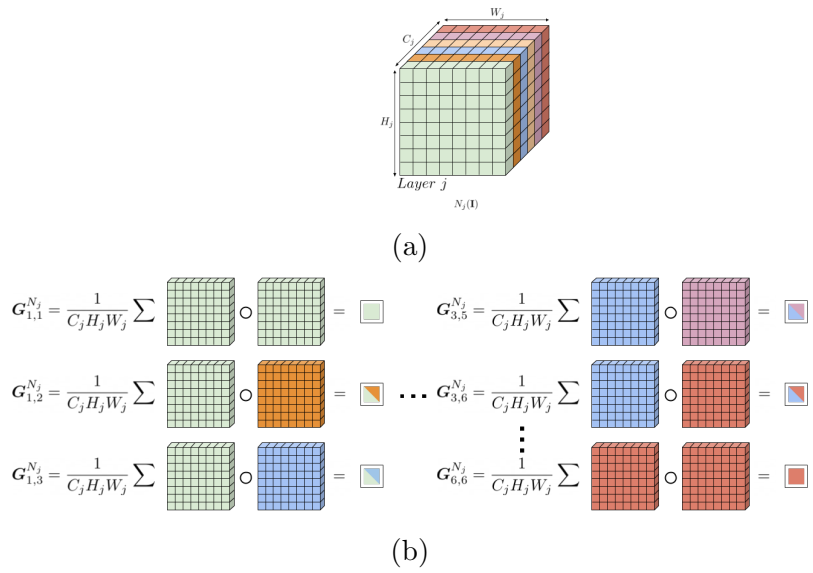
Given an input image $\mathbf{I}$ and a neural network N_e made of e hidden layers which can be interpreted as the functional composition of e functions L_j followed by a final mapping $\overline{L}$ that depends on the task: $N_e = \overline{L} \circ L_e \circ ... \circ L_1$. Let $N_j(\mathbf{I})$ be the feature maps of the j-th layer of the network N_e for the input $\mathbf{I}$, with $j \leq e$, which can be seen as a matrix of shape $C_j \times H_j \times W_j$ resulted from the application of the functional composition of the first j functions $L_1, \cdots, L_j$ of the network N_e such that $N_j = L_j \circ ... \circ L_1$. For the layer j-th the Gram matrix $\mathbf{G}^{N_j}$ can be defined as a matrix of shape $C_j \times C_j$ whose elements are given by:

$$G^{N_j}_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} N_j(\mathbf{I})_{c,h,w} N_j(\mathbf{I})_{c',h,w} \tag{2.1}$$

where c and c' are the indices of the Gram matrix, and both vary in the range $[1, C_j]$. In the Figure 2.2 is reported an illustration of how the Gram matrix is computed from a feature maps $N_j(\mathbf{I})$.

$\mathbf{G}^{N_j}$ captures which features tend to activate together, therefore $\mathbf{G}^{N_j}$ is excepted to be a symmetric matrix with a negligible diagonal representing the correlation between a filter and itself. For the purpose of the proposed approach, only the lower triangular matrix of $\mathbf{G}^{N_j}$ can be considered, which once flattened results in a feature vector $\mathbf{x}_j$ of size $1 \times [(C_j(C_j+1)/2) - C_j]$. For each layer j of the VGG16, in the Table 2.1 is reported the resulting feature vector dimension representing the intra-layer correlation.

The Gram matrix is invariant with respect to the size of the input image $\mathbf{I}$ since its dimension depends only on the number of filter C_j of the j-th layer, and it can be computed efficiently if $N_j(\mathbf{I})$ is reshaped into a matrix $\boldsymbol{A}$ of size $C_j \times (H_j \times W_j)$ so that $\boldsymbol{G}^{N_j} = \boldsymbol{A}\boldsymbol{A}^T / C_j H_j W_j$. In Figure 2.3 is reported a schematic overview of the Gram matrix efficient computation.



(a)

(b)

Figure 2.2: Schematic view of Gram matrix computation. In (**a**) is reported the feature maps of the j-th layer. (**b**) illustrate, for some of the indices of the Gram matrix, how they are computed. The symbol $\circ$ refers to the element-wise matrix product.

2.3.2 Anomaly Detection

The proposed anomaly detection technique is inspired by the approach presented in [96]. The degree of abnormality, that is the presence of distortions in the input image, is computed by measuring the similarity between the aforementioned intra-layer correlation representation of a given image, and a reference dictionary $\boldsymbol{\mathcal{W}}$ of Gram matrices computed from a database of pristine images. The similarity is measured through the Euclidean distances between the feature vector $\mathbf{x}$ extracted from a given image $\mathbf{I}$ and the feature vectors of the dictionary $\boldsymbol{\mathcal{W}}$. The final *abnormality score* is the sum of the average distances and α times the standard deviation of the distances. It is defined as follow:

$$Abnormality\ score = \frac{1}{D}\sum_{d=1}^{D} dist(\mathbf{x}, \mathbf{w}_d) + \alpha\sqrt{\frac{1}{D}\sum_{d=1}^{D}\left(dist(\mathbf{x}, \mathbf{w}_d) - \frac{1}{D}\sum_{d=1}^{D} dist(\mathbf{x}, \mathbf{w}_d)\right)} \tag{2.2}$$

where D is the number of words in the reference dictionary $\boldsymbol{\mathcal{W}} = \{\mathbf{w}_1, ..., \mathbf{w}_D\}$, and $dist(\mathbf{x}, \mathbf{w}_d)$ is the Euclidean distance between the feature vector $\mathbf{x}$ representing the input image $\mathbf{I}$ and the words of the dictionary $\mathbf{w}_d$.

The dictionary is built from a subset of pristine images $\boldsymbol{\mathcal{P}} = \{\mathbf{I}_1, .., \mathbf{I}_D\}$: for each image $\mathbf{I}_d$ the Gram matrix is computed respect a given jth layer of the network N_e and keep only the flattened lower triangular matrix of $\mathbf{G}^{N_j}$ of shape $(C_j(C_j+1)/2) - C_j$. The dimension of the feature vector is then reduced to M with $M < (C_j(C_j+1)/2) - C_j$ by applying the Principal Component Analysis (PCA) [130] to all the feature vector computed from $\boldsymbol{\mathcal{P}}$. M represents the number of principal components such that a given percentage of the

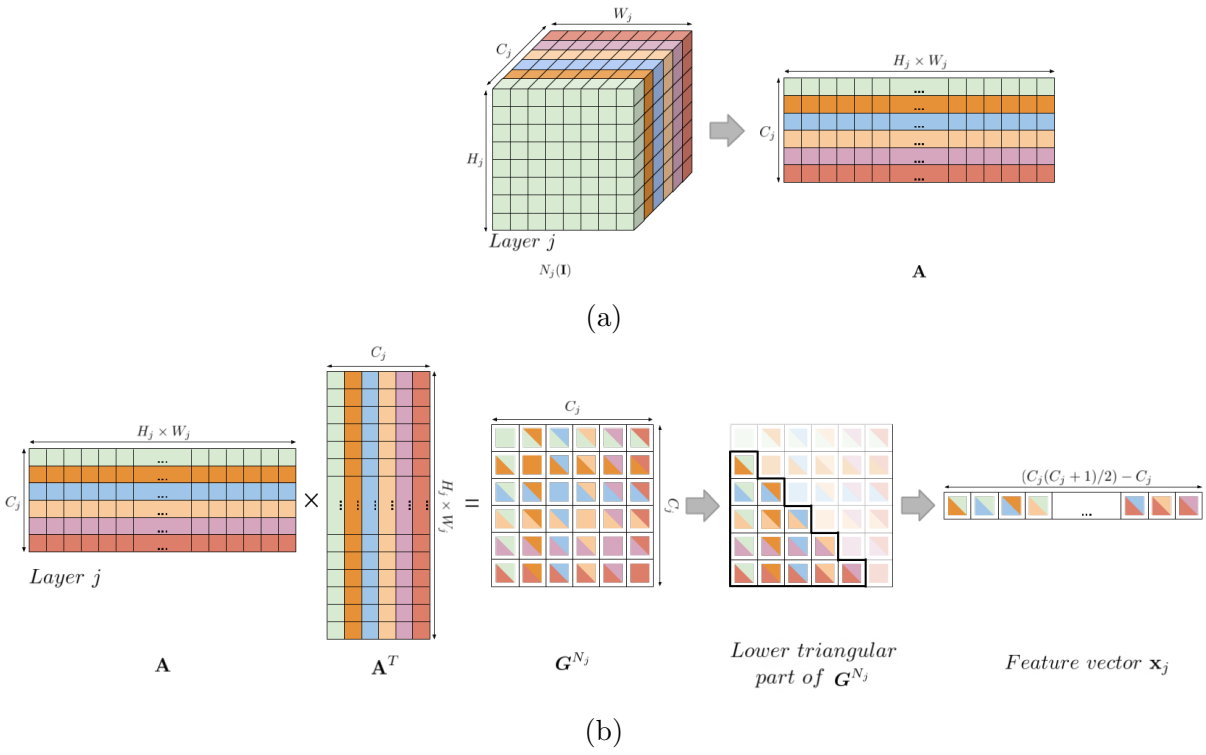


Figure 2.3: Visual overview of the Gram matrix efficient computation and feature vector extrapolation. In (**a**) is shown how the activation volume is reshaped to efficiently compute the Gram matrix in (**b**) to finally compute the feature vector that represents the intra-layer correlation.

data variance is retained.

Finally, all the reduced features vectors of $\boldsymbol{\mathcal{P}}$ are grouped into clusters using Mean Shift [23], a centroid-based algorithm which use kernel density estimation to iteratively move the sample to the nearest local maxima of the probability density with respect to the input samples. It only has a parameter named bandwidth which represents the width of the kernel, and the output are the best k clusters corresponding to k centroids. The advantage of using Mean Shift as clustering algorithm is that it does not require to explicitly define the number of clusters k in advance. The number of clusters is instead variable and depends solely on the data and the width of the kernel chosen. A centroid is the most representative point within the cluster, and in this case, is the mean position of all the elements of the cluster. The considered dictionary $\boldsymbol{\mathcal{W}}$ is therefore composed by the coordinates of the center of each cluster found.

A schematic view of the dictionary creation pipeline is presented in the Figure 2.4.

Table 2.1: Summary of the VGG16 [114] architecture alongside the resulting feature vector dimension with respect to the intra-layer correlation.

Block	Layer (Name)	Layer (Type)	Kernel Size	# Filters	Feature Vector Dimension
1	conv1_1	Convolutional	3×3	64	2016
	conv1_2	Convolutional	3×3	64	2016
	Max_Pooling	Pooling	-	-	-
2	conv2_1	Convolutional	3×3	128	8128
	conv2_2	Convolutional	3×3	128	8128
	Max_Pooling	Pooling	-	-	-
3	conv3_1	Convolutional	3×3	256	32,640
	conv3_2	Convolutional	3×3	256	32,640
	conv3_3	Convolutional	3×3	256	32,640
	Max_Pooling	Pooling	-	-	-
4	conv4_1	Convolutional	3×3	512	130,816
	conv4_2	Convolutional	3×3	512	130,816
	conv4_3	Convolutional	3×3	512	130,816
	Max_Pooling	Pooling	-	-	-
5	conv5_1	Convolutional	3×3	512	130,816
	conv5_2	Convolutional	3×3	512	130,816
	conv5_3	Convolutional	3×3	512	130,816
	Max_Pooling	Pooling	-	-	-
6	fc6	Dense			
7	fc7	Dense			

2.3.3 Combining Method

The final quality score is the combination of factors: (1) the arithmetic mean of the CNNs correlations and (2) the *abnormality score*.

To make these two factors comparable they are scaled[26]. Moreover, the correlation of the *abnormality score* is flipped by computing 1—*abnormality score*. Then, with the purpose of having better interpretability, final score is divided by 2 and multiplied by 100 in order to have values that ideally range between 0 and 100 as the MOS.

Since the anomaly detection method relies on the concept of distance from a dictionary of pristine images, it implies that the closer the input image is to this dictionary, the less abnormal is the considered picture. On the contrary, the farther the input representation is from the dictionary, the higher is the *abnormality score*. The output of the anomaly detection method is therefore negatively correlated with the Mean Opinion Scores.

On the other hand, it is reasonable to believe that a quality artefact in an input image of a CNN, may alter the activation maps, hence it affects negatively the average intra-layer correlation: the more corrupted (less qualitative appealing) is the input image, the less correlated are the layers of the network. It reflects that the average of the intra-layer correlation is positively correlated with the Mean Opinion Scores.

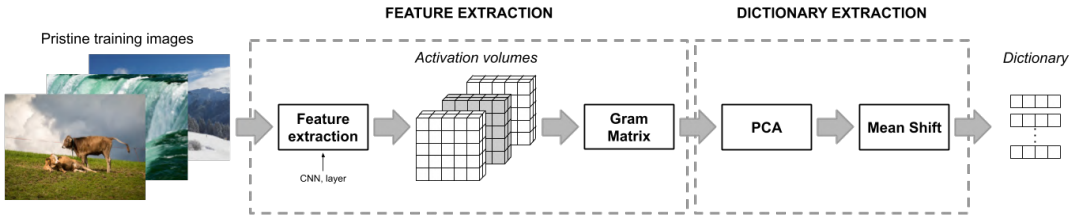


Figure 2.4: Overview of the creation of the dictionary for the degree of abnormality computation. The activation volumes of a given CNN are extracted and then the Gram matrix is computed to get the intra-layer correlation for a subset of pristine images. Subsequently dimensionality reduction is applied through Principal Component Analysis (PCA). Finally Mean Shift algorithm is performed to compute clusters on which centroids are then extracted as entry of the dictionary.

2.4 Experiments

In this section, first the databases taken into account for the experiments are described, followed by the metrics used for the evaluation of the proposed approach and then the implementation details of the method.

2.4.1 Database for Image Quality Assessment with Real Distorted Images

There are several databases for image quality assessment with real distorted images. For the experiments we considered one of the most used dataset in the field of IQA which is the LIVE in the Wild Image Quality Challenge Database [40] (LIVE-itW) and two recent large scale IQA databases namely: Smartphone Photography Attribute and Quality database (SPAQ) [34] and the KonIQ-10k [47] (KONIQ).

The task of gathering opinion scores on the perceived quality of images is highly subjective and subtle, for all the previously mentioned databases, therefore, every worker was first provided with detailed instructions to help them assimilate the task. Since LIVE-itW and KonIQ-10 rely on crowdsourcing frameworks, a selection of participants based on the worker's reliability was applied. Also, during and after the process of data annotation several filtering steps were implemented to ensure an acceptable level of quality for the resulting Mean Opinion Scores. To further validate the credibility of the collected scores, the authors of the three databases, also reports the mean inter-group agreement as the mean agreement between the MOS values of non-overlapping random groups of users in terms of Spearman's rank ordered correlation. In particular, the LIVE-itW database reaches a value of 0.9896 while KonIQ-10 and SPAQ measure a correlation of 0.973 and 0.923 respectively.

The LIVE-itW database [40] is a collection of 1162 authentically distorted images captured from many diverse mobile devices. The images have been evaluated by over

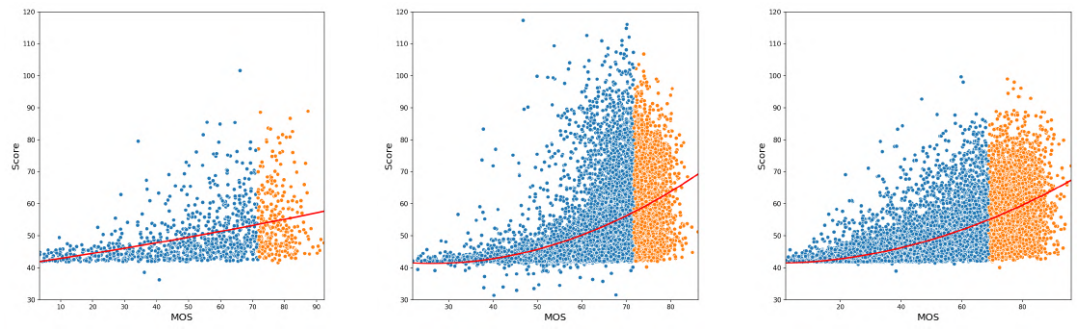
8100 human observers through Amazon Mechanical Turk and each image has been viewed and rated on a continuous quality scale by an average of 175 unique subjects. The Mean opinion scores ranges from 0 to 100.

The KONIQ database [47] contains 10,073 images selected from the Yahoo Flickr Creative Commons 100 Million dataset (YFCC100M) [121]. Each image has a total of 120 reliable quality ratings obtained by crowd-sourcing, performed by 1459 subjects. The collected values of the Mean Opinion scores belongs to the interval $[0, 5]$, for readability, they are scaled in the range $[0, 100]$ to be coherent with the other datasets.

Finally, the SPAQ database [34] consists of 11,125 pictures taken by 66 smartphones. The images have been labeled through a subjective test, in a well-controlled laboratory environment, with a total o 600 subjects. Each image has been rated at least 15 times. The MOS varies between 0 and 100.

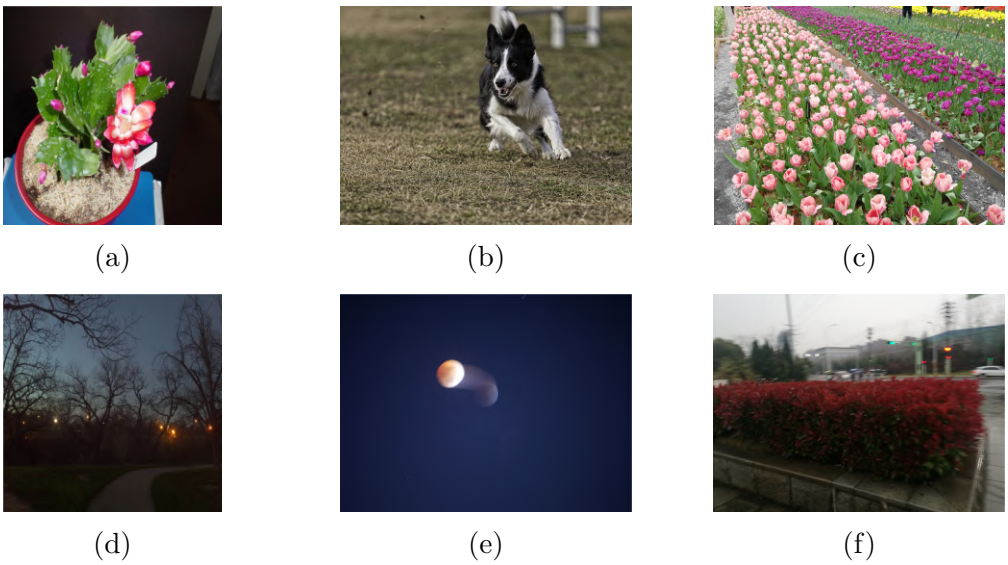
The datasets for image quality assessment with real distorted images are divided into good quality images and bad quality images according to the MOS. As a threshold, for each dataset, it has been decided to take the 75th percentile with respect to the MOS distribution. Therefore, images having a MOS over the 75th percentile are labelled as good quality images, while the remaining ones are labeled as poor quality images. These thresholds are 71.71 and 71.74 for the KONIQ and LIVE-itW respectively while for the SPAQ the 75th percentile respect the MOS is 68.82.

In the Figure 2.5 is reported the distributions of the Mean Opinion Scores of the three datasets where it can be seen that KONIQ and LIVE-itW seam to distribute similarly with unimodal behaviour, while the SPAQ's distribution of the MOS appear to be bimodal. An overview of database properties is provided in Table 2.2 and samples images alongside MOS are in Figure 2.6.



(a) (b) (c)

Figure 2.5: Estimated density distribution of Mean Opinion Scores for the three datasets: (**a**) LIVE in the Wild Image Quality Challenge Database (LIVE-itW), (**b**) KonIQ-10k (KONIQ), and (**c**) Smartphone Photography Attribute and Quality database (SPAQ). The bars represents the normalized histogram, the blue line is the estimated density distribution while in red line is the 75th percentile respect the Mean Opinion Score (MOS).



(a) (b) (c)

(d) (e) (f)

Figure 2.6: Sample images form the three Image Quality Assessment (IQA) databases: (**a**) and (**d**) images from the LIVE-itW with a MOS of 78.81 and 43.67 respectively, (**b**) and (**e**) KONIQ's pictures with a MOS of 71.46 and 43.36; finally (**c**) and (**f**) photos form SPAQ whit a MOS of 75.43 and 33.0 respectively.

Table 2.2: Overview of databases used for image quality assessment with real distorted images.

Database	**Year**	**No. of Images**	**Rating per Images**	**Environment**	**Inter-Group Agreement**	**MOS 75th Percentile**
LIVE-itW [40]	2015	1162	175	crowd-sourcing	0.9896	71.74
KONIQ [47]	2018	10,073	120	crowd-sourcing	0.9730	71.71
SPAQ [34]	2020	11,125	15	lab	0.9230	68.82

2.4.2 Database for Image Quality Assessment with Real Synthetic Distortion

As pristine collection of images, has been taken into account a large scale database for image quality assessment with synthetic distortion named Konstanz Artificially Distorted Image quality Set (KADIS700k) [74]. For the purpose or the presented method, only the non-distorted pictures have been considered. The dataset is composed of 140,000 pristine images collected from Pixabay.com, a website for sharing photos and videos. According to the image quality guidelines of Pixabay.com, users must upload pictures that have a well defined subject, clear focus, and compelling colours. Also, images with chromatic aberration, JPEG compression artefacts, image noise, and unintentional blurriness are not accepted. Moreover, authors of KADIS700k have collected up to twenty independent votes by Pixabay users claiming that the quality rating process provides a reasonable

indication that the released images are pristine. After selecting 654,706 images with resolutions higher than 1500×1200 they randomly sampled 140,000 images as pristine reference images. Under the aforementioned conditions is believed that KADIS700k can model the concept of pristine images necessary to build the proposed system.

An example of the pictures present in the dataset is reported in Figure 2.7.

(a) (b) (c)

Figure 2.7: Sample images form the KADIS700k databases. (**a**-**c**) random photos from KADIS700k.

2.4.3 Experimental Setup

The most commonly used metrics to evaluate the performance of no-reference image quality assessment methods are respectively the Pearson's Linear Correlation Coefficient (PLCC) and the Spearman's Rank-order Correlation Coefficient (SROCC). Both are used to compare the scores predicted by the models and the subjective opinion scores provided by the dataset (for a detailed description see in the section of this thesis A.1 in the appendix).

The figures of merit used to assess the ability of the methods to discriminate good vs. bad quality images were: the area under a receiver operating characteristic (ROC) curve, abbreviated as AUC, and the area under the precision-recall curve (AUPR). In particular the AUC measures the overall performance of a binary classifier, and it ranges between 0.5 and 1.0, where the minimum value corresponds to the performance of a random classifier while the maximum value represents the oracle. Meanwhile, the AUPR reflects the trade-off between precision and recall as the threshold varies.

2.4.4 Implementation Details

The proposed method is trained and tested using Python. The training process consists of extracting the information of intra-layer correlation, from a subset of pristine images of KADIS700k [74], and compute the dictionary $\boldsymbol{\mathcal{W}}$ of k centroids trough Mean Shift. Then the average intra-layer correlation and the degree of abnormality is computed from a different subset of pristine images of KADIS700k [74] in order to save the minimum and maximum value of both to lately perform min-max scaling in order to merge them. The parameter α for the anomaly detection method has been chosen performing a parameter

tuning over a subset of synthetically distorted images for KADIS700k, resulting in a value of 2.

For the intra-layer correlation, have been used the ImageNet [108] pre-trained VGG16 provided by the Torchvision package of the PyTorch framework [102]. The model was trained scaling the input images into the range of [0, 1] and then normalized using mean $[0.485, 0.456, 0.406]$ and standard deviation $[0.229, 0.224, 0.225]$. As preprocessing images were first cropped such that the resulting size was evenly distributed between 8% and 100% of the original image area and then a random aspect ratio between 3/4 and 4/3 of the original aspect ratio was applied. The resulting crops were finally resized to 224 $\times$ 224. As layer for the representation have been used the output of the first convolutional layer of the second convolutional stage ($conv2\_1$) of the VGG16, which consists of 128 filters, resulting in a feature vector of shape 8192. Moreover, the input of the VGG16 images are scaled so the smaller edge results of 512 pixels and normalized according to the mean value and standard deviation of ImageNet.

For the training phase, 10,000 pristine images from KADIS700k are selected, the Gram matrix is computed for each image. Subsequently, PCA is applied with a percentage of retained variance of 97%. Finally, Mean Shift algorithm is used with a flat kernel to find the representative centroids and build the dictionary for the computation of the *abnormality score*. The bandwidth for the Mean Shift was empirically estimated as the average pairwise distance between the samples that are in the same cluster applying the Nearest Neighbors algorithm. The resulting bandwidth is 1.804.

To perform the min-max scaling on both the *abnormality score* and the average value of the intra-layer correlation, 1000 pristine images from KADIS700k are randomly selected and different from the previous ones and the minimum and maximum values for the two metrics are collected.

The proposed method is compared against three different no-reference benchmark algorithms for the IQA, one opinion unaware, which does not require MOS during the training phase named ILNIQE [136] and two opinion aware methods which require the mean opinion scores of the images on which they are trained, namely CORNIA [133] and DIIVINE [94]. For these methods, their original implementations alongside the saved models released from the authors are taken. For this reason, these methods are executed a single time.

2.5 Results

We compared the average performance in terms of SROCC, PLCC, AUC, and AUPR on the three considered databases (LIVE-itW, KONIQ, and SPAQ) across 100 iterations. Here below we report the overall (average) and by dataset performance.

2.5.1 Average Performance

Table 2.3 reports the average performance over the three databases of the proposed opinion-unware method and state-of the-art methods, both opinion unware and opinion aware. This table permits to have a quick look of the best performing method whatever is the database considered. The proposed method is on average the first in terms of SROCC, AUC and AUPR against all the methods. In terms of PLCC, the proposed method is the second best opinion-unware method.The PLCC gap is believed to be caused by the fact that the proposed method is not trained using Mean Opinion Score hence it is not forced to have a linear relationship with the target distribution. Moreover, as discussed in Section 2.4.3, SROCC is a more suitable metric than PLCC for the evaluation of ordinal variables.

Table 2.3: Pearson's Linear Correlation Coefficient (PLCC), Spearman's Rank-order Correlation Coefficient (SROCC) area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (AUPR) of the proposed solution respect the three state of the art methods (ILNIQUE, CORNIA, DIIVINE). In each column, the best values are marked in boldface.

Method	**SROCC**	**PLCC**	**AUC**	**AUPR**
ILNIQUE † [136]	0.5596 ± 0.0000	0.5489 ± 0.0000	0.7327 ± 0.0000	0.4507 ± 0.0000
CORNIA * [133]	0.5638 ± 0.0000	**0.5859 ± 0.0000**	0.7329 ± 0.0000	0.4336 ± 0.0000
DIIVINE * [94]	0.5109 ± 0.0000	0.4870 ± 0.0000	0.7431 ± 0.0000	0.4501 ± 0.0000
Proposed method †	**0.5989 ± 0.0021**	0.4823 ± 0.0036	**0.7659 ± 0.0009**	**0.4710 ± 0.0006**

† opinion-unaware methods, * opinion-aware methods.

In terms of correlation, DIIVINE appears to be the least correlated according to both SROCC and PLCC. ILNIQE and CORNIA are very close in terms of SROCC but they perform worse with respect to the proposed method, while on the PLCC, CORNIA results to be the most effective method, followed by ILNIQE.

Regarding the capability of the methods to discriminate good quality images from the bad ones, in terms of AUC, the CORNIA, and ILNIQE methods have similar performance, followed by the DIIVINE method which performs slightly better, and finally, by the proposed method that achieves the best performance. Regarding the AUPR, the behavior is very similar to the AUC excepts for the DIIVINE method which results to be better than CORNIA.

2.5.2 Single Dataset Performance

The average results obtained on the three datasets show the generalization skills of the proposed algorithm. In Table 2.4 is reported the performance of the methods under examination for each dataset.

On the LIVE-itW database the performance highlights that the proposed method performs better in terms of SROCC with respect to all methods, while it is the best in

Table 2.4: SROCC, PLCC, AUC, and AUPR of the proposed solution compared with the three from state of the art methods (ILNIQUE, CORNIA, DIIVINE) for each datasets taken into account (LIVE-itW, KONIQ, and SPAQ). For each database, in each column, the best values are marked in boldface.

Evaluation Dataset	Method	SROCC	PLCC	AUC	AUPR
LIVE-itW	ILNIQUE †	0.4516 ± 0.0000	**0.4968 ± 0.0000**	0.6610 ± 0.0000	0.3501 ± 0.0000
	CORNIA *	0.4307 ± 0.0000	0.4819 ± 0.0000	0.6598 ± 0.0000	0.3710 ± 0.0000
	DIIVINE *	0.4599 ± 0.0000	0.4504 ± 0.0000	**0.7127 ± 0.0000**	**0.4140 ± 0.0000**
	Proposed method †	**0.4933 ± 0.0018**	0.4058 ± 0.0028	0.7031 ± 0.0012	0.4112 ± 0.0007
KONIQ	ILNIQUE †	0.5130 ± 0.0000	0.5026 ± 0.0000	0.7214 ± 0.0000	**0.4480 ± 0.0000**
	CORNIA *	0.5510 ± 0.0000	**0.5654 ± 0.0000**	0.7294 ± 0.0000	0.4218 ± 0.0000
	DIIVINE *	0.4734 ± 0.0000	0.4322 ± 0.0000	0.7177 ± 0.0000	0.4329 ± 0.0000
	Proposed method †	**0.5741 ± 0.0042**	0.4256 ± 0.0042	**0.7446 ± 0.0011**	0.4049 ± 0.0007
SPAQ	ILNIQUE †	0.7142 ± 0.0000	0.6473 ± 0.0000	0.8156 ± 0.0000	0.5539 ± 0.0000
	CORNIA *	0.7096 ± 0.0000	**0.7103 ± 0.0000**	0.8094 ± 0.0000	0.5080 ± 0.0000
	DIIVINE *	0.5993 ± 0.0000	0.5784 ± 0.0000	0.7989 ± 0.0000	0.5035 ± 0.0000
	Proposed method †	**0.7292 ± 0.0002**	0.6155 ± 0.0038	**0.8501 ± 0.0003**	**0.5970 ± 0.0003**

† opinion-unaware methods, * opinion-aware methods.

terms of AUC and AUPR against all the opnion-unware methods but the second best (with a negligible difference of about 0.006 on average) with respect to the DIIVINE, which is an opinion-aware method. In terms of PLCC, ILNIQE is the best then followed by CORNIA, DIIVINE and the proposed method. The LIVE-itW is the database on which all the methods perform worse with respect to the others datasets. This can be caused by the dimension of the dataset, which is one-tenth of the others.

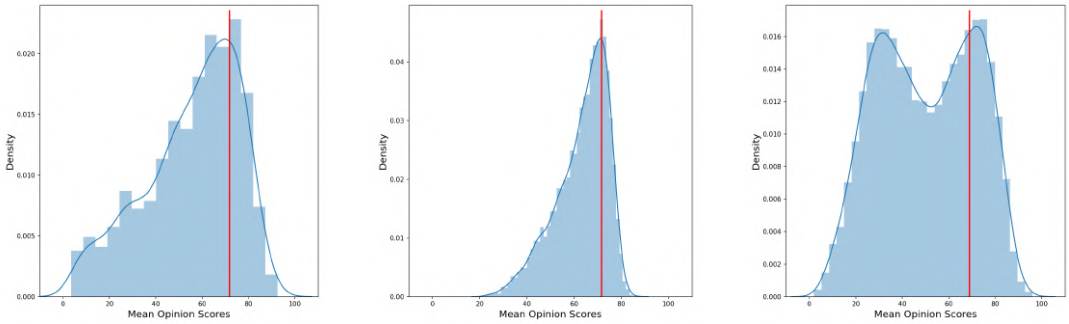
On the KONIQ database, the proposed method performs better in terms of SROCC and AUC with respect to all methods. Surprisingly, ILNIQE reaches the top performance on AUPR even if on the other measures is not highly competitive. The best in terms of PLCC is the CORNIA. Even in this case, performance of all methods, regardless the metric, are quite low.

Finally, on the SPAQ database, the overall behavior follows the one presented in the average performance (cf. Table 2.3): the proposed method is the best in terms of SROCC, AUC, and AUPR while is not competitive in terms of PLCC.

Summing up, the proposed method is first in terms or SROCC against all the methods over the three considered databases. It is the best opinion-unaware method in terms of AUC while it is the second best against an opinion-aware method only on the LIVE-itW dataset with a difference of 0.096. Since the proposed method does not force any kind of linear relationship between the output and the data distribution, it is not competitive in terms of PLCC: it is the lowest performing method except for the SPAQ database were it is placed third. Considering the AUPR metric, except for the KONIQ database, the proposed method is the best of all methods on SPAQ and the second best on the LIVE-itW database of a small amount (0.0028).

Figure 2.8 shows, for each of the databases, the predicted quality score distributions with respect to the ground truth MOS. While in the Figure 2.9 are reported two pictures

from the three dataset LIVE-itW, KONIQ, and SPAQ, alongside the predicted image quality scores and the mean opinion scores.



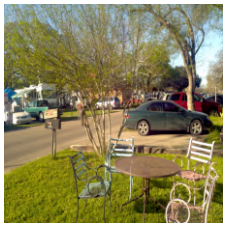
(a) (b) (c)

Figure 2.8: Scatter plots of the predicted quality scores respect MOS for the three considered datasets: (**a**) LIVE-itW, (**b**) KONIQ, and (**c**) SPAQ. In red is depicted the second-order interpolation line. The points in blue belongs to the bad quality images (MOS $< 75°$ percentile of the MOS distribution for that dataset) while the orange ones refer to the good quality images.

2.5.3 Ablation Study

In this section the focus is on the contribution given by the anomaly detection method to the solely average of the intra-layer correlation. Table 2.5 shows that the intra-layer correlation achieves competitive results in terms of SROCC, AUC, and AUPR while is weaker with respect to the PLCC. On the other hand, the output from the anomaly detection method tends to be less effective on the four metrics. Although, the *abnormality score* tends to be more competitive on the PLCC except on the KONIQ database on which the PLCC scores is very low.

Even if the average intra-layer correlation alone shows interesting performance, combined with the *abnormality score* results in an increase of the PLCC while maintaining similar performance on the other metrics.



(a)
QS: 67.94 MOS: 75.69

(b)
QS: 63.91 MOS: 71.27

(c)
QS: 73.37 MOS: 82.67

(d)
QS 42.71 MOS: 16.13

(e)
QS: 44.33 MOS: 52.64

(f)
QS: 42.79 MOS: 42.79

Figure 2.9: Examples of predicted quality score (QS) alongside the mean opinion score (MOS). First column images (**a**,**d**) belong to LIVE-itW database, second column picures (**b**,**e**) are from KONIQ dataset while the last column photos (**c**,**f**) belong to the SPAQ collection.

Table 2.5: SROCC, PLCC, AUC and AUPR of the proposed solution alongside the average intra-layer correlation (MEAN_GM) and the *abnormality score* for the databases LIVE-itW, KONIQ and SPAQ. For each database, in each column, the best values are marked in boldface.

Dataset	**Method**	**SROCC**	**PLCC**	**AUC**	**AUPR**
LIVE-itW	MEAN_GM	0.4930 ± 0.0000	0.3520 ± 0.0000	**0.7043 ± 0.0000**	0.4068 ± 0.0000
	ABNORMALITY SCORE	0.4599 ± 0.0075	0.3609 ± 0.0266	0.6826 ± 0.0045	0.3966 ± 0.0035
	Proposed method	**0.4933 ± 0.0018**	**0.4058 ± 0.0028**	0.7031 ± 0.0012	**0.4112 ± 0.0007**
KONIQ	MEAN_GM	0.5512 ± 0.0000	0.3050 ± 0.0000	0.7379 ± 0.0000	0.3998 ± 0.0000
	ABNORMALITY SCORE	0.4873 ± 0.0220	0.1343 ± 0.0335	0.7169 ± 0.0102	**0.4380 ± 0.0082**
	Proposed method	**0.5741 ± 0.0042**	**0.4256 ± 0.0042**	**0.7446 ± 0.0011**	0.4049 ± 0.0007
SPAQ	MEAN_GM	0.7281 ± 0.0000	0.5227 ± 0.0000	**0.8505 ± 0.0000**	0.5934 ± 0.0000
	ABNORMALITY SCORE	0.6590 ± 0.0152	0.4387 ± 0.0526	0.8037 ± 0.0104	0.5525 ± 0.0100
	Proposed method	**0.7292 ± 0.0002**	**0.6155 ± 0.0038**	0.8501 ± 0.0003	**0.5970 ± 0.0003**
AVERAGE	MEAN_GM	0.5908 ± 0.0000	0.3932 ± 0.0000	0.7642 ± 0.0000	0.4667 ± 0.0000
	ABNORMALITY SCORE	0.5354 ± 0.0149	0.3113 ± 0.0376	0.7344 ± 0.0084	0.4624 ± 0.0072
	Proposed method	**0.5989 ± 0.0021**	**0.4823 ± 0.0036**	**0.7659 ± 0.0009**	**0.4710 ± 0.0006**

Chapter 3

Modeling image aesthetics through aesthetics-related attributes

3.1 Introduction

Easy access to a camera and the consequent nearly effortless task of taking photos has made shooting a picture similar to natural action. We took photos in every moment of our days, for example to remind something or to capture events. Despite cameras are becoming increasingly sophisticated and smart, it is not rare to shoot images that are not pleasing in terms of aesthetics. Given the exponential growth of the number of images taken and stored, selecting pleasing images has become a tedious and boring task. Being able to automatically distinguish good aesthetics images from bad ones can help various types of applications, such as automatic photo album creation, media storage techniques and so on.

This thesis proposes a method based on a MLP that, from features extracted by an ImageNet pretrained CNN, predicts eleven aesthetics-related attributes. Then a SVR [104] is trained to predict image aesthetics on the basis of the aesthetics-related attributes computed in the previous stage.

The main contributions of this thesis in the direction of predicting the image aesthetics through aesthetics-related attributes are the following:

- To introduce an aesthetics quality estimation method that relies on the prediction of aesthetics-related attributes.

- To show that predicting the aesthetics of an image is more accurate through aesthetics-related attributes rather than modeling only the aesthetics or jointly the aesthetics-related attributes and the global aesthetics.

- To demonstrate the effectiveness of the proposed approach which outperformed the state of the art of about 5.5% in terms of Spearman's Rank-order Correlation Coefficient (SROCC) over the "Aesthetics with Attributes Database" (AADB).

3.2 Related work

Automatic aesthetics assessment of images is usually treated as a classification or regression task based on ratings provided by human annotators [6]. In recent years, there has been a lot of research effort and various approaches have been proposed.

Datta et al. [28] carefully selected 56 hand-crafted visual features based on standard photography and visual design rules to discriminate between aesthetically pleasing and displeasing images.

Dhar et al. [31] proposed a method for predicting image interestingness by exploiting high-level describable image attributes divided into three categories: compositional (image layout or configuration), content (objects or scene types depicted) and sky-illumination (natural lighting conditions).

With the availability of more labeled data the trend has been shifting from methods based on hand crafted features to deep learning. Recent works have both been focused on sophisticated training loss [63, 111, 100, 20] and more powerful features [91, 106, 48].

Given the importance of photography rules and aesthetics attributes, Kong et al. have collected the "Aesthetics with Attributes Database", or AADB [63]. This collection includes images that have been rated by several human observers in terms of both global aesthetics and visual aesthetics-related attributes. They proposed a Convolutional Neural Network (CNN) architecture to jointly predict semantic photo content, global aesthetics and aesthetics-related attributes.

Malu et al. [91] proposed a multi-task network based on features extracted from a ResNet-50 [44]. To better encapsulate the information from the ResNet-50 they extracted 16 rectified convolution maps from the ReLU output of the 16 residual blocks of the ResNet-50. The proposed architecture is used to predict eight aesthetics attributes alongside the global aesthetics score.

In [111] the authors explored the relationship between aesthetics score and aesthetics attributes introducing the PI-DCNN: a ResNet optimized over three different loss functions: regression, ranking and a privileged information loss which rely on some domain knowledge and additional information between attributes and aesthetics.

In [100] Pan et al. exploited the feature extracted from a ResNet-50, and proposed a multi-task neural network to predict both the aesthetics score and the attributes. Different from the other works, they proposed a framework in which the network is trained in an adversarial manner: the discriminator distinguishes the predictions given by the proposed multi-task network from the real labels.

Chen [20] proposed a different training framework based on data covariance learning to improve performance of baseline architectures: the method proved that training an architecture modeling the data uncertainty is more effective than training with the mean squared error.

In [106] authors combined low-resolution, semantically strong features with the high-resolution, semantically weak features from the EfficientNets B4 [119] to predict simultaneously 8 aesthetics tags and the global aesthetics score.

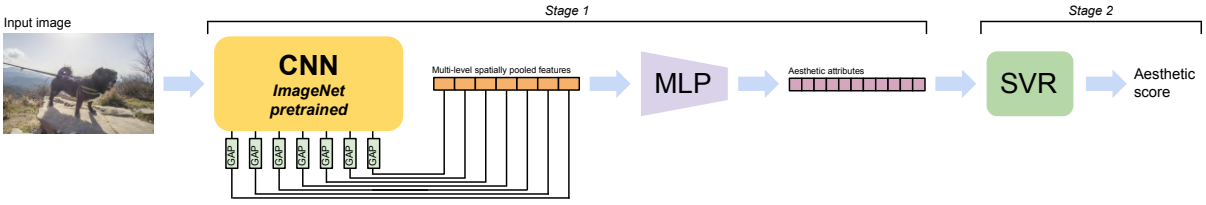


Figure 3.1: The proposed method. Given an input image, a multi-level spatially pooled features set is extracted from a Convolutional Neural Network pretrained on ImageNet. This feature set is then fed to a Multi Layer Perceptron to predict image aesthetics-related attributes. Finally a Support Vector Regression machine is used to estimate the image aesthetics score starting from the aesthetics-related attributes.

3.3 Proposed method

Figure 3.1 shows the pipeline of the proposed method. The first stage is a MLP that predicts the eleven aesthetics-related attributes of an input image on the basis of MLSP features extracted from an ImageNet pretrained network. The second stage is based on a SVR that takes the attributes predicted in the first stage as input and it estimates the global image aesthetics score. The design choices have been driven by the following considerations.

MLSP features demonstrated to be very effective for image aesthetics prediction [48]. The main idea was to create a feature vector that encodes information from multiple levels of a CNN. This is achieved by concatenating Global Average Pooled (GAP) activations from fixed blocks of a given CNN trained for image classification.

The use of aesthetics-related attributes for the estimation of image aesthetics has been investigated in previous works [28][31]. Datta et al. [28] proposed several visual attributes and studied the correlation between those properties and the aesthetics score. Some examples of attributes are: light exposure, colorfulness, depth of field and rule of thirds. They also proposed a Support Vector Classification machine that uses 15 visual attributes for the classification of high and low rated photographs in terms of interestingness.

Dhar et al. [31] proposed an aesthetics estimation method that from low-level features (e.g. Color spatial distribution map, Spatial Pyramid of shape features etc.) predicts aesthetics-related attributes such as presence of a salient objects, opposing colors, presence of people and clear skies.

The proposed method takes inspiration from the paper by Hosu et al. [48] for what concerns the use of MLSP features and from the paper by Dhar et al. [31] for what concerns the use of aesthetics-related attributes to predict image aesthetics.

To assess the effectiveness we experimented different CNN architectures (ResNet-50, EfficientNets B4, Inception-v3, InceptionResNet-v2) pretrained on ImageNet [5, 108] from which the MLSP features are extracted. The proposed method is also compared with two variants: a single-task MLP trained to predict solely the aesthetics score and a multi-task MLP trained jointly over the eleven aesthetics attributes and the aesthetics score. The

Table 3.1: Correlation between aesthetics properties and the aesthetics scores.

Property	**SROCC**
Balacing elements	0.3830
Color harmony	0.6227
Content	0.7279
Depth of field	0.5098
Light	0.6221
Motion blur	0.2204
Object	0.6415
Repetition	0.1023
Rule of thirds	0.3892
Symmetry	0.1063
Vivid color	0.6161
All of the above (SVR)	0.9374

performance was measured in terms of Spearman's Rank-order Correlation Coefficient (see also Appendix A.1 for further details and formulae).

3.4 Experiments

3.4.1 Dataset

The proposed method is trained and tested on the AADB [63], a database composed of 10,000 images. Each image of the database has the aesthetics rating and the assessment of eleven aesthetics-related attributes provided by five different subjects. The images are divided into training (8,500), validation (500) and testing sets (1,000), and they were collected from the Flickr website and curated manually. With the help of professional photographers the authors has selected eleven attributes that are closely related to image aesthetics judgements: *interesting_content, object_emphasis, good_lighting, color_harmony, vivid_color, shallow_depth_of_field, motion_blur, rule_of_thirds, balancing_element, repetition,* and *symmetry.*
To gather the data, authors ask qualified Amazon Mechanical Turk (AMT) workers to rate "positive" if an attribute conveyed by the image can enhance the image aesthetics level, or "negative" if the attribute degrades image aesthetics. According to the authors, the default value was "null", meaning that the attribute does not affect image aesthetics. The collected labels were then translated into real values encoding 'positive' as 1, 'negative' as -1 and 'null' as 0. For each image, the attribute score is the average over all the users judgements. Figure 3.2 reports, for each of the eleven aesthetics attributes, the

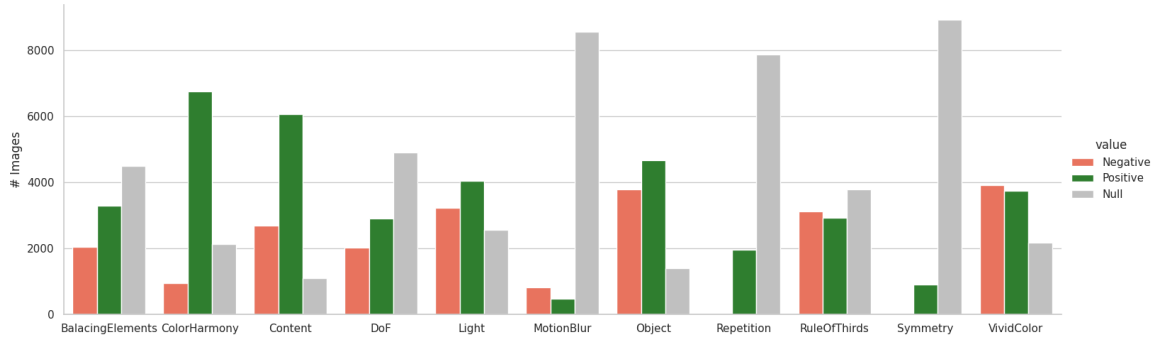


Figure 3.2: Value distribution for each of the eleven aesthetics attributes. Null values are those which have a mean score of 0. Positive values are those images with on average more positive labels than negative, vice-versa for the Negative.

distribution of the mean values to underlying imbalances of the ground truths: attributes like motion blur, symmetry or repetition contains many "null" values.

For the aesthetics score, AMT workers were allowed to express their judgement on a scale from 1 to 5. For each image, the aesthetics score is the average over all the users judgements. The aesthetics score was further normalized in order to fit a range of [0, 1]. Figure 3.3 depicts the distribution of the aesthetics scores.

In order to highlight the intrinsic power of the aesthetics-related attributes, table 3.1 reports the SROCC between the values of each attribute and the overall aesthetics score. Color harmony, Content, Light, Object, Vivid color correlate more than others with the image aesthetics with an SROCC higher than 0.6. To better highlight the prediction power of the aesthetics-related attributes, a SVR is trained for image aesthetics score estimation based on the human ground truth. On the whole all the attributes achieve an SROCC value higher than 0.9. We also investigated the correlation between the SROCC of each of these attributes and the percentage of null values in the ground truth and it has been found a SROCC of -0.9182. This suggests that the poor correlation of some attributes with the aesthetics score is more likely due to the null values rather than the expressiveness of the attribute itself.

3.4.2 Experimental Setup

The models have been developed in PyTorch and trained on an Nvidia GTX 1070 GPU. The MLP is composed of three stacked linear layers with ReLu activations. The training has been done with a batch size of 16 and for a maximum of 100 epochs adopting the early stop technique with patience of 6 epochs over the average of the SROCC with respect to the validation set. As optimizer was used Adam [59] with a learning rate of 1.5e-05. The first two layers of the MLP were trained with a dropout probability of 0.5.

For the feature extraction part, as in [48], it has been decided to extract and store

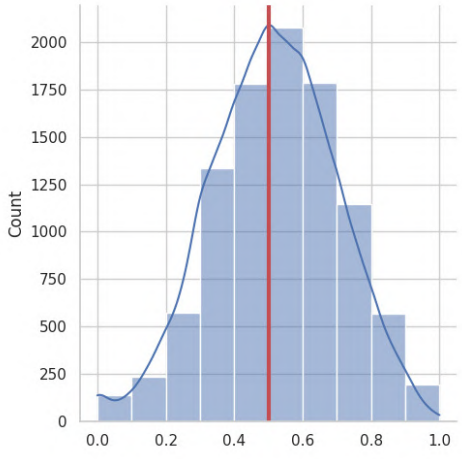


Figure 3.3: Score distribution of the AADB database. The red line indicates the 0.5 value.

from images having a different resolution, fixed sized narrow MLSP features of dimension $(1 \times 1 \times b)$ where b is the number of kernels from which features are computed. To extract these featuress it has been adopted the Global Average Pooling layer (GAP) over selected activation blocks output: for the EfficientNets B4, Inception-v3 and InceptionResNet-v2 it has been decided to select the same block as done by the original works ([106, 48]). Note that for the EfficientNets B4 authors extract features from given blocks in a different way, while for the ResNet-50 have been selected all the five convolutional blocks. There are 6 blocks in EfficientNets B4 (3,056 kernels), 11 blocks in Inception-v3 (10,048 kernels), 43 in InceptionResNet-v2 (49,248 kernels) and 5 blocks in ResNet-50 (3,904 kernels).

The most commonly used metric to evaluate the performance of automatic image aesthetics assessment is the SROCC. It is used to compare the scores predicted by the models and the subjective opinion scores provided by the dataset and it evaluates the monotonic relationship between two continuous or ordinal variables. The SROCC operates on the rank of the data points ignoring the relative distances between them.

3.5 Results

As stated before, to better understand the contribution given by the proposed method, the proposed pipeline is compared against two different alternatives which are common in the state of the art: a single-task Neural Network [6, 48] trained to predict the aesthetics score directly from the image and a multi-task Neural Network [63, 100] trained to predict at the same time the eleven aesthetics attributes and the aesthetics score. Table 3.2 reports the mean of SROCC achieved on 10 repetitions of the experiments. The table reports, for each of CNN archictecture experimented, the proposed approach and the two variants mentioned above. The table also reports results taken from the most recent state-of-the-

Table 3.2: Spearman's Rank-order Correlation Coefficient (SROCC) between the predicted image aesthetics quality and the ground truth. (* srocc are taken from the authors pubblication)

Name (base architecture)	**Architecture type**	SROCC
Kong et al. (Alexnet) [63]	Multi-Task CNN	0.6782*
Malu et al. (Resnet-50) [91]	Multi-Task CNN	0.6890*
PI-DCNN (Resnet-50) [111]	Multi-Task CNN	0.7051*
Chen (Resnet-50) [20]	Multi-Task CNN	0.7080*
Pan et al.(ResNet-50) [100]	Multi-Task CNN	0.7041*
Reddy et al. (Efficientnet_b4) [106]	Multi-Task CNN	0.7059*
EfficientNets_B4	Single-Task MLP	0.7281 ± 0.0138
	Multi-Task MLP	0.7219 ± 0.0039
	SVR over MLP's tag prediction	**0.7454** ± 0.0033
ResNet-50	Single-Task MLP	0.7083 ± 0.0067
	Multi-Task MLP	0.7194 ± 0.0060
	SVR over MLP's tag prediction	**0.7384** ± 0.0023
Inception-v3	Single-Task MLP	0.7242 ± 0.0065
	Multi-Task MLP	0.7197 ± 0.0029
	SVR over MLP's tag prediction	**0.7354** ± 0.0025
InceptionResNet-v2	Single-Task MLP	0.7316 ± 0.0029
	Multi-Task MLP	0.7308 ± 0.0036
	SVR over MLP's tag prediction	**0.7429** ± 0.0015

art approaches which jointly predict aesthetics-related attributes and image aesthetics.

Overall, independently from the architectural choice, the proposed method outperforms the state of the art of about 5.5% in terms of SROCC. The improvement of the proposed method with respect to the other two alternatives is of about 2.4%. The enhancement given by the proposed method confirms that predicting image aesthetics through the estimation of aesthetics-related attributes is more effective than a multi task CNN. Moreover, it is more effective predicting the aesthetics score on the basis of aesthetics-related attributes rather than predicting the aesthetics score along with attributes or solely the aesthetics score. Figure 3.4 shows an example of the predicted images attributes with respect to the ground truth values.

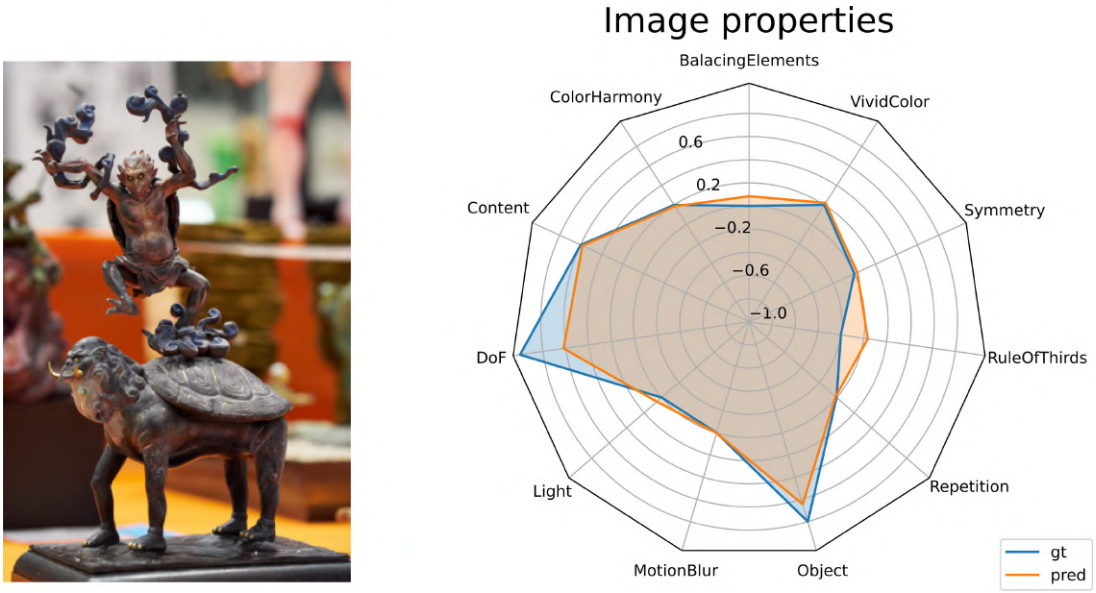


Figure 3.4: Example of predicted aesthetics-related attributes (orange line) with respect to the ground truth (blue line).

Chapter 4

Incorporating Composition and Style Knowledge into a CNN for Image Aesthetic Assessment

4.1 Introduction

Aesthetics of an image is defined as the measure or appreciation of the beauty of an image. Image aesthetics is a subjective property that depends on the viewer's preferences, experience, and skills as a photographer. Despite this, the occurrence of specific factors or patterns makes an image objectively more appealing than others. Researchers have in fact found that aesthetics can be influenced by several factors including lighting [35], color scheme [109], contrast [53], composition [78], semantic photo content [16, 84], and image styles [97, 62].

The semantic content of a photo is a key aspect in the evaluation of aesthetic quality: (i) psychology research shows that certain kinds of content are more attractive than others [36]; (ii) professional photographers choose different photographic techniques and have different aesthetic criteria in mind when shooting different types of contents [28, 84].

In the same way, image styles such as "Long Exposure", "Macro", "Bokeh" and others, or image geometric composition rules such as "Rule of Thirds", "Curved" and others, influence the aesthetic quality of an image [97, 62]. Figure 4.1 shows an example of a high-quality image and a low-quality image. In this example, the image on the top has been rated on average by a group of humans with high level of aesthetics. This is likely due to nice attributes such as, good lighting and harmonious color combinations, which make the image attractive. In contrast, the image below has a low aesthetic level rate which likely is due to low light and dull colors.

Literature reports mostly on three different aesthetics recognition tasks: high vs. low aesthetic quality [82, 55, 87, 78], aesthetic score regression [6, 48, 111, 100], and aesthetic score distribution prediction [118, 141, 19]. Whatever is the recognition task, most of researchers do not explicitly model the aforementioned factors that influence image

Figure 4.1: Two images with high and low esthetics from the AADB database [63]. The left image has a high aesthetic likely thanks to good lighting and harmonious color combinations, while the image with a low aesthetic has low light and dull colors.

aesthetics and indeed they prefer holistic approaches [28, 84, 93, 48]. Besides, methods explicitly modeling aesthetic attributes try to learn a universal model to be applied to images with different aesthetic attributes [100, 111, 69]. However, images that share the same aesthetic attribute or particular combinations of attributes can also have different levels of aesthetics. It is therefore necessary to learn aesthetic prediction models that specialize for each type of attribute and at the same time are able to consider the correlation between several attributes.

Here we describe a new method that aims to resolve the aforementioned limitations. The proposed method models the aesthetics of the image by explicitly taking into account image semantic content, style and composition. In particular, it exploits privileged side information related to aesthetic attributes, that already demonstrated to be effective for improving the aesthetic modeling of the image [91, 100]. Instead of having a single aesthetic estimator to govern the images with whatever aesthetic attributes you want, this thesis moves to the opposite in which each image based on its attributes has an *ad hoc* estimator of aesthetic quality.

To better exploit side privileged information given by image style and composition attributes, the training stage of the proposed method is multi-stage. The first training stage involves a Multi Layer Perceptron (MLP) network (the AttributeNet) which is specially to recognize image style and composition. This network takes as input semantic features extracted by a pre-trained network (the Backbone). The second training stage concerns a hypernetwork (HyperNet).

The latter exploits the attributes prior encoded into the embedding generated by the AttributeNet to predict the parameters of the target network dedicated to aesthetic estimation (AestheticNet). The adoption of the attribute-conditioned hypernetwork, therefore, determines attribute-specific aesthetic estimators. The HyperNet is trained using the Earth Mover's Distance (EMD) as a loss function to better learn the distribution of user judgments attributed to each image. This strategy allows to model the consensus and the diversity of opinions among the annotators and consequently to improve the effectiveness of the proposed method.

Given a test image, the proposed method predicts image style and composition as well

as the aesthetic score distribution.

To summarize, the contribution of this thesis with respct to the image aesthetic assessment are the following.

- To present a deep learning-based method that not only estimate aesthetics in terms of score distribution but also determine the style and composition of the input image.
- To propose a hypernetwork which adaptively generates the aesthetic quality prediction parameters basing on aesthetic attributes. The proposed method predicts image aesthetics in a content and attribute aware manner, therefore it is not limited to a holistic evaluation of the aesthetic quality.
- To carry out comprehensive experiments for unified aesthetic prediction tasks: aesthetic classification, aesthetic regression, and aesthetic label distribution. For all of these tasks, the proposed method achieves higher performance than state-of-the-art approaches on three common benchmark datasets (AADB [63], AVA [95], and Photo.net [28]).

4.2 Related work

In this section, first, it is reviewed the relevant literature related to image aesthetic quality assessment. Then it is highlighted the differences between the proposed method and similar existing methods.

4.2.1 Image aesthetic quality assessment

From the seminal work of Datta *et al.* [28] many research efforts have been made, and various methods have been proposed for estimating the aesthetics of images [30]. Several papers proposed the use of hand-crafted features to encapsulate both aspects of human perception and photographic rules. For example, Datta *et al.* [28] carefully selected 56 hand-crafted visual features based on standard photographic rules (such as rule of thirds, colorfulness, or saturation) to discriminate between aesthetically pleasing and displeasing images. Luo *et al.* [84] extracted features encoding photographic rules, e.g. composition, lighting, and color arrangement, to evaluate aesthetics in different ways based on the photo content. Zhang *et al.* [137] modeled image aesthetics by focusing on the image composition which is modeled using graphlets small-sized connected graphs.

However, methods based on hand-crafted features can only achieve limited success [30]: (i) hand-crafted features can not exhaustively model the variations of photographic rules between different categories of images; (ii) hand-crafted features are heuristics, and so it is challenging to mathematically model some photographic rules. Based on the previous considerations and thanks to the availability of more labeled data, the trend has shifted from hand-crafted feature-based methods to deep learning methods [6, 46, 143].

RAPID [82] is a double-column network that captures both local and global information of images for discriminating low and high aesthetics. Given that one patch may not well represent the fine-grained information in the entire image, Lu *et al.* [83] extended RAPID by proposing Deep Multi-patch Aggregation Network (DMA-Net). In DMA-Net, an input image is represented by a bag of random cropped patches. The proposed layers, namely the statistics and sorting layers, enabled the integration of multiple input patches. Given that DMA-Net failed to encode the global layout of the image, Ma *et al.* presented the Adaptive Layout-Aware Multi-Patch Convolutional Neural Network (A-Lamp CNN) [87]. This method images of arbitrary size as input and learns from both fine-grain details and holistic image layout simultaneously. It consists of two subnets, i.e. a Multi-Patch subnet which is very similar to DMA-Net and a Layout-Aware subnet consisting of an object-based attribute graph. Multi-Net Adaptive Spatial Pooling ConvNet (MNA-CNN) [90] is trained and tested on images at their original sizes and aspect ratios. It computed aesthetics by combining multi-level features and scene information. Chen *et al.* [19] designed the Adaptive Fractional Dilated Convolution (AFDC) that, similarly to MNA-CNN, avoids altering original image aspect ratio and composition. RGNet [78] builds a region graph to represent the visual elements and their spatial layout in the image, and then performs reasoning on the graph to uncover the mutual dependencies of the local regions. Gated Peripheral-Foveal Convolutional Neural Network (GPF-CNN) [141] is a deep architecture designed to: encode the global image composition; extract the fine-grained details from aesthetic-relevant regions.

Professional photographers adopt different photographic techniques and have various aesthetic criteria depending on the portrait content. Therefore, Kao *et al.* [55] proposed a Multi-task Convolutional Neural Network (MTCNN). This model aims to simultaneously estimate the semantic content and the aesthetic class of an image.

Neural IMage Assessment (NIMA) [118] replaced the classification layer of a pre-trained ImageNet CNN with a fully-connected regression head that predicts the distributions of ratings per image. The squared Earth Mover's Distance (EMD) has been employed as the loss function. In this thesis, the network for aesthetic score distribution is optimized by minimizing EMD loss, and approach similar to the one proposed in [118]. Multi-level Spatially Pooled activation blocks (MLSP) [48] exploited a transfer learning strategy that uses features extracted from a pre-trained ImageNet CNN.

Some recent works regard multi-modal aesthetic evaluation models that leverage visual information along with user comments. The latter encodes high-level semantic information and are relevant for aesthetic decisions [143, 46, 140].

4.2.2 Correlation with existing methods

Several state-of-the-art methods are similar to the one proposed in this thesis in that they use multiple attributes describing the aesthetic or artistic aspect of a photo for aesthetic evaluation [63, 69, 37, 111, 100].

Leonardi *et al.* [69] and Gao *et al.* [37] used attributes as mid-level representation

to estimate the aesthetics of the image. Thus, the errors on the attributes might be propagated to the assessed aesthetics. Pan *et al.* [100] proposed a multi-task deep network to learn the aesthetic score and aesthetic attributes simultaneously. Attributes were used as additional information for the learning paradigm called Learning Using Privileged Information (LUPI) [123]. Besides, adversarial learning was introduced to capture the correlation between the aesthetic score and attributes. Shu *et al.* [111] also exploited LUPI by proposing Deep Convolutional Neural Network with Privileged Information (PI-DCNN): a novel method exploring photo attributes as privileged information for photo aesthetic assessment.

There are three major differences between the previous works and the one proposed in this thesis:

- First, previous methods are limited to the datasets annotated with aesthetic attributes, namely AVA or AADB. In contrast, in the proposed method, side information about composition and style is learned from specially designed datasets. The proposed method can therefore generalize to a larger number of aesthetic attributes.

- Second, the previous methods learn a universal model of aesthetics that depends indiscriminately on the aesthetic attributes. In contrast, the proposed method learns aesthetic models that are dependent on the different aesthetic attributes present within the image and their correlation.

- Finally, Earth Mover's Distance (EMD) is used as a loss function to better learn the distribution of user judgments attributed to each image. This strategy allows to model the consensus and the diversity of opinions among the annotators and consequently to improve the effectiveness of the proposed method.

4.3 Proposed method

Given an input image $\mathbf{X}$, the goal of the proposed method is to estimate both the aesthetic score distribution $\hat{\mathbf{q}}$ and the presence of a set of aesthetic-related attributes $\hat{\mathbf{y}}$ by using the network f parametrized with θ^*:

$$\mathbf{X} \xrightarrow{f(\theta^*)} (\hat{\mathbf{q}}, \hat{\mathbf{y}}). \tag{4.1}$$

More specifically, $f \leftarrow (f_s, f_t)$ consists of two networks. The network f_s handles the *side* information regarding the aesthetic-related attributes and produces the final output $\hat{\mathbf{y}}$ and the embedding $\mathbf{e}_s$. The latter is exploited by an *attribute-conditioned* hypernetwork $\hat{\theta}_t = h(\mathbf{e}_s; \theta_h^*)$ that adaptively generate the parameters $\hat{\theta}_t$ of the network f_t. Such a network f_t carries out the *main* task of aesthetic assessment. Hence, by using the attribute-conditioned hypernetwork the aesthetic assessment task is subordinated to that of attribute estimation.

$\theta^* \leftarrow (\theta_h^*, \theta_s^*, \theta_b^*)$ is the set of learned parameters of the proposed method. Unlike the θ_b^* parameters which belong to a pre-trained backbone, the others are learned for the tasks. Similar to [4], a two-step optimization procedure is adopted to introduce attribute-constraint into the hypernetwork.

The first step regards the training of the parameters θ_s of the network f_s. Let $\mathcal{D}_s = \{(\mathbf{X}_s^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^N$ denotes the training set of N training samples. Each training sample consists of a color image $\mathbf{X}_s^{(i)} \in \mathbb{R}^d$ and K aesthetic-related attributes $\mathbf{y}^{(i)} \in \mathbb{R}^K$. Given the training set $\mathcal{D}_s$, the goal is to learn a network $f_s : \mathbb{R}^d \rightarrow \mathbb{R}^K$, which predicts whether an attribute occurs or not into the input image:

$$\mathcal{L}_{side}\{f_s(\mathbf{e}_b; \theta_s), \mathbf{y}) \mid \theta_s \in \Theta_s\}, \tag{4.2}$$

where $\theta_s \in \Theta_s$ are the learnable real-valued parameters and $\mathbf{e}_b = b(\mathbf{X}_s; \theta_b^{*(L)})$ is the embedding corresponding to the activations of layer L of the pre-trained backbone b given the input $\mathbf{X}_s$.

The second training concerns the hypernetwork. Instead of directly learning the θ_t parameters of the network f_t, the hypernetwork is trained to learn the parameters θ_h of a metamodel h. The output of this metamodel is $\hat{\theta}_t$. The network h can therefore be thought of as a generator of parameters to obtain attribute-specific aesthetic estimators. Let $\mathcal{D}_t = (\mathbf{X}_t^{(i)}, \mathbf{q}^{(i)})_{i=1}^N$ denotes the training set of N training samples. Each training sample consists of a color image $\mathbf{X}_t^{(i)} \in \mathbb{R}^d$ and a distribution of aesthetics ratings $\mathbf{q}^{(i)} = [q_{s_1}, q_{s_2}, ..., q_{s_B}]$. Where s_j is the j-th score bucket, B is the total number of score buckets, and q_{s_j} denotes the number of voters that give the discrete score s_j to the image. Given the training set $\mathcal{D}_t$, the goal is to learn the parameters θ_h of the metamodel to generate the parameters for the network $f_t : \mathbb{R}^d \rightarrow \mathbb{R}^B$, which predicts the aesthetic score distribution $\hat{\mathbf{q}}$:

$$\mathcal{L}_{task}\{f_t(\mathbf{e}_b; h(\mathbf{e}_s; \theta_h)), \mathbf{q}) \mid \theta_h \in \Theta_h\}, \tag{4.3}$$

where $\mathbf{e}_b$ is the same embedding as in Eq. 4.2, $\mathbf{e}_s = f_s(\mathbf{e}_b; \theta_s^{*(M)})$ is the attribute-conditioned embedding obtained from the M-th layer of the pre-trained f_s given the input $\mathbf{e}_b$, and $\theta_h \in \Theta_h$ are the learnable real-valued parameters.

4.3.1 Proposed network architectures

The proposed architecture includes four different networks trained using a multi-stage approach: the Backbone, the AttributeNet, the HyperNet and the AestheticNet. The overall architecture of the model is shown in Figure 4.2.

The Backbone is an ImageNet [29] pre-trained neural network that outputs multi-level features used as inputs of the AestheticNet and the AttributeNet. The latter is a Multi Layer Perceptron (MLP) network specially trained for image style and composition recognition. The features obtained from the previously trained AttributeNet are then used as HyperNet inputs. It is a metamodel dedicated to calculating the weights and distortions

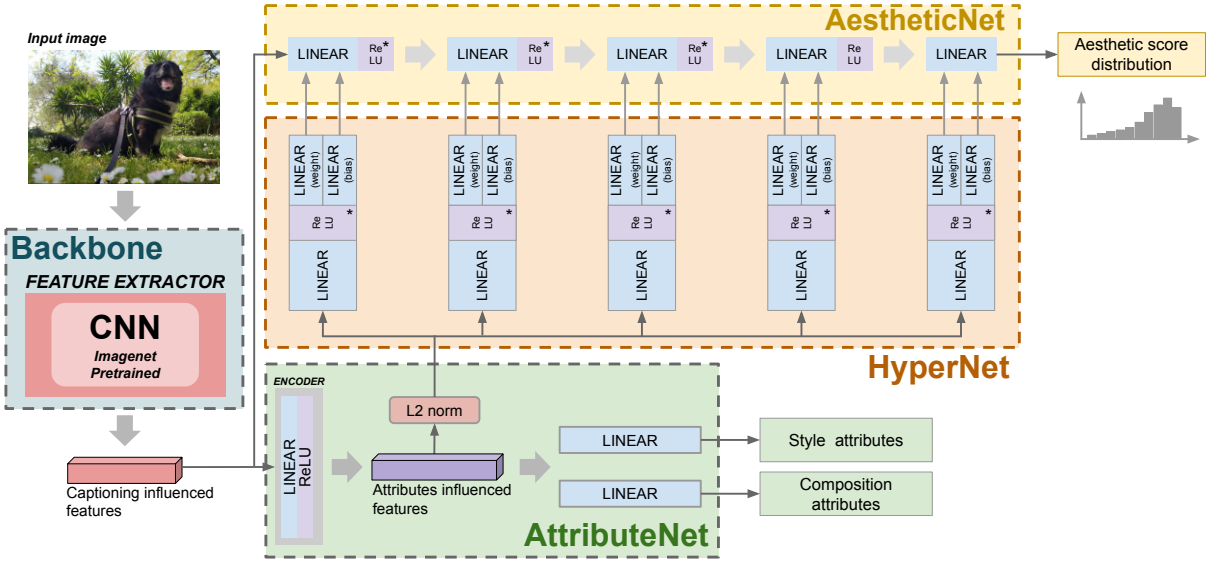


Figure 4.2: The proposed method is composed of four main parts: the Backbone, the AttributeNet, the HyperNet and the AestheticNet. The input image is first fed to the Backbone to extract a feature set that encodes the content of the image. Then, this feature set is fed to AttributeNet. The goal of the AttributeNet is to predict aesthetics-related attributes and influence the input of the HyperNet. The HyperNetwork aims to predict the weights and the biases of the AestheticNet. Finally, the AestheticNet infers the aesthetic score distribution of the input image over the content related feature set with the weights and the biases predicted by the HyperNet. *Trained with dropout

of the AestheticNet. Therefore, the HyperNet is trained to allow the AestheticNet to predict aesthetics scores for input images in close agreement with human judgments.

Backbone

Many of the earlier approaches to image aesthetics assessment extract features from warped and/or cropped input images [6, 118, 111]. A shortcoming of these methods is that they alter the composition of the image and the aspect ratio of the objects. Thus, they may harm the task of aesthetics assessment.

On the other hand, previous works have shown the effectiveness of multi-level features to predict perceptual judgements, either for image quality assessment [75] or image aesthetics assessment [48, 106, 69].

For the previous reasons, the Backbone network $b : \mathbf{X} \rightarrow \mathbf{e}_b$ encodes an input image, at the original size, $\mathbf{X} \in \mathbb{R}^d$ into a Multi-Level Spatially Pooled (MLSP) embedding vector $\mathbf{e}_b \in \mathbb{R}^D$. The resulting embedding vector encodes information at multiple levels of abstraction: from low- to high-level features. This goal is achieved by stacking activations from L layers of a given pre-trained CNN. As the spatial resolution of the different activation maps varies, a Global Average Pooling (GAP) is adopted to squeeze the spatial

dimensions into a channel activation vector. Therefore, the size of the MLSP embedding vector depends solely on the number of channels in each layer. This last aspect allows processing images at full resolution without the need to resize or crop them.

To summarize, the Backbone network $b(\mathbf{X}; \theta_b^{*(L)})$ outputs the embedding $\mathbf{e}_b$ corresponding to the activations of L layers given the input $\mathbf{X}$.

An ImageNet-pretrained EfficientNet-B4 [119] is exploited as Backbone network b: a very efficient yet effective model not only on ImageNet but also on transfer learning datasets. Recently, it demonstrates its effectiveness for image aesthetic assessment [106, 69]. Following [106], the activations of the MBConv blocks [119] having numbers $L = \{15, 21, 25, 29, 31\}$ are considered. Given the input image $\mathbf{X}$ with shape $h \times w \times 3$, the resulting activation maps have shape: $\frac{h}{16} \times \frac{w}{16} \times 112$, $\frac{h}{16} \times \frac{w}{16} \times 160$, $\frac{h}{32} \times \frac{w}{32} \times 272$, $\frac{h}{32} \times \frac{w}{32} \times 272$, and $\frac{h}{32} \times \frac{w}{32} \times 448$. The previous 5 activation maps are spatially narrowed using the GAP and stacked on the channel dimension, thus obtaining a fixed sized narrow MLSP embedding vector of shape $1 \times 1 \times 1264$.

AttributeNet

The AttributeNet $f_s : \mathbf{e}_b \rightarrow \hat{\mathbf{y}}$ with $\hat{\mathbf{y}} \in \mathbb{R}^K$ is a Multi Layer Perceptron (MLP) that aims to categorize the backbone embedding $\mathbf{e}_b$ with respect to K aesthetic-related attributes. More specifically, $\hat{\mathbf{y}} \leftarrow (\hat{\mathbf{y}}_v, \hat{\mathbf{y}}_c)$, where $\hat{\mathbf{y}}_v \in \mathbb{R}^{K_v}$ is the set of image styles and $\hat{\mathbf{y}}_c \in \mathbb{R}^{K_c}$ is the set of composition rules. Therefore, the MLP performs two tasks simultaneously and consists of three linear blocks. The first block is a linear layer with ReLU that transforms the embedding vector $\mathbf{e}_b$ into an embedding vector, $\mathbf{e}_s$:

$$\mathbf{e}_s = \text{ReLU}(\mathbf{W}_s^\top \mathbf{e}_b + \mathbf{b}_s). \tag{4.4}$$

Given that $\mathbf{e}_s$ is shared between the two tasks, it intrinsically encodes the style and composition as well as their relationships. The second and third blocks are independent linear layers categorizing the input image into style and composition:

$$\hat{\mathbf{y}}_v = \mathbf{W}_v^\top \mathbf{e}_s + \mathbf{b}_v, \tag{4.5}$$

$$\hat{\mathbf{y}}_c = \mathbf{W}_c^\top \mathbf{e}_s + \mathbf{b}_c, \tag{4.6}$$

where $\mathbf{W}_v$ and $\mathbf{b}_v$ are the parameters for predicting the style $\hat{\mathbf{y}}_v$, and $\mathbf{W}_c$ and $\mathbf{b}_c$ are the parameters for predicting the composition $\hat{\mathbf{y}}_c$.

AestheticNet

The AestheticNet $f_t(\mathbf{e}_b; \hat{\theta}_t)$ aims to predict the aesthetic score distribution $\hat{\mathbf{q}}$ given the embedding vector $\mathbf{e}_b$ produced by the Backbone. It is a MLP composed of M linear layers whose parameters $\hat{\theta}_t = \{(\hat{\mathbf{W}}_1, \hat{\mathbf{b}}_1), (\hat{\mathbf{W}}_2, \hat{\mathbf{b}}_2), ..., (\hat{\mathbf{W}}_M, \hat{\mathbf{b}}_M)\}$ are computed by the

HyperNet h:

$$\mathbf{x}_1 = \text{ReLU}(\hat{\mathbf{W}}_1^\top \mathbf{e}_b + \hat{\mathbf{b}}_1), \tag{4.7}$$

$$\mathbf{x}_i = \text{ReLU}(\hat{\mathbf{W}}_i^\top \mathbf{x}_{i-1} + \hat{\mathbf{b}}_i), \quad \text{with} \quad i = 2, ..., M-1 \tag{4.8}$$

$$\hat{\mathbf{q}} = \hat{\mathbf{W}}_M^\top \mathbf{x}_{i-1} + \hat{\mathbf{b}}_M. \tag{4.9}$$

In general, the input of a linear layer is $\mathbf{x}_{in} \in \mathbb{R}^{N_{in}}$, and the corresponding output is $\mathbf{x}_{out} \in \mathbb{R}^{N_{out}}$. Thus, the weights of the layer corresponds to $\mathbf{W} \in \mathbb{R}^{N_{in} \times N_{out}}$ and the bias are equal to $\mathbf{b} \in \mathbb{R}^{N_{out}}$. The number of parameters in a linear layer is $N_{in}N_{out}$ for $\mathbf{w}$ and N_{out} for $\mathbf{b}$.

HyperNet

The HyperNet $h(\mathbf{e}_s; \theta_h^*)$ is the metamodel that generates the parameters $\hat{\theta}_t$ for the M layers of the AestheticNet. Therefore, the HyperNet is composed of M HyperNet Blocks (HBs), $h = \{HB_1, HB_2, ..., HB_M\}$.

Each HB is composed of a linear layer that reduces the size of the embedding $\mathbf{e}_s \in \mathbb{R}^D$ to a size of $d \mid d \ll D$:

$$\mathbf{e}_r = \text{ReLU}(\mathbf{W}_i^{r\top} \mathbf{e}_s + \mathbf{b}_i^r), \tag{4.10}$$

where ReLU is the activation function, $\mathbf{W}_i^r \in \mathbb{R}^{D \times d}$ are the learned weights, and $\mathbf{b}_i^r \in \mathbb{R}^d$ are the learned bias. Two linear layers are then dedicated to the estimation of the parameters, $\mathbf{W}_i$ and $\mathbf{b}_i$ of the i-th layer of the AestheticNet:

$$\hat{\mathbf{W}}_i = \mathbf{W}_i^{w\top} \mathbf{e}_r + \mathbf{b}_i^w, \tag{4.11}$$

$$\hat{\mathbf{b}}_i = \mathbf{W}_i^{b\top} \mathbf{e}_r + \mathbf{b}_i^b. \tag{4.12}$$

Given the generated weights $\hat{\mathbf{W}}_i \in \mathbb{R}^{N_{in} \times N_{out}}$, the learned parameters of the linear layer are $\mathbf{W}_i^w \in \mathbb{R}^{d \times N_{in} N_{out}}$ and $\mathbf{b}_i^w \in \mathbb{R}^{N_{in} N_{out}}$, respectively. Instead, for the generated bias $\hat{\mathbf{b}}_i \in \mathbb{R}^{N_{out}}$, the learned parameters of the linear layer are $\mathbf{W}_i^b \in \mathbb{R}^{d \times N_{out}}$ and $\mathbf{b}_i^b \in \mathbb{R}^{N_{out}}$.

4.4 Experiments

In this section, first the considered datasets are described, then the evaluation metrics used to estimate the performance and the training procedure of the proposed method are described.

4.4.1 Datasets

Datasets for aesthetic-related attribute recognition

Most previous methods that exploit the relationship between attributes and aesthetic quality rely on the use of the aesthetic attributes provided for the AVA and AADB

datasets [37, 111]. Although the two sets of attributes are valuable as they span the traditional photographic principles of color, lighting, focus, and composition, they have some drawbacks. First, they are not exchangeable or can be merged because their annotation has a different meaning: in the AVA dataset, annotations are binary values that indicate the occurrence of the attribute in the image; for AADB, the annotation of each attribute can assume continuous values between -1 and 1, where "positive" and "negative" indicates that the occurrence of the attribute improves or degrades the aesthetic level of the image, respectively. Second, only a subset of the AVA images (about 4.44%) provides the style annotations. Third, the number of attributes in the two sets is limited. Few attributes categorize images for composition style (e.g., Rule of Thirds and symmetry). Although the role of emotion in the aesthetic experience is proven [25], no attributes are specifying what emotions the image content conveys.

Based on the above considerations, a different and broader set of attributes is considered in this work. Attributes were chosen by taking into account not only the composition but also the style of an image. Among these, there are the optical techniques used during the shot (such as bokeh effect and depth-of-Field), the genre of the image content (e.g., horror or romantic), the atmospheric light conditions (such as hazy or sunny), finally, the mood aroused by the image (e.g., serene).

In this thesis it has been used KU-PCP [67] and FlickrStyle [56] dataset. The set $\mathcal{D}_t$ is used to train the AttributeNet and it is composed by the two datasets mentioned above. Table 4.1 compares the lists of attributes present in the AADB, the AVA datasets and the list of the 29 attributes taken from the KU-PCP and FlickrStyle datasets.

FlickrStyle: The FlickrStyle dataset [56] is a collection of 80,000 photographs gathered from the Flickr website annotated with 20 curated style labels. These can be categorized into:

- *Atmosphere:* Hazy, Sunny
- *Color:* Bright, Pastel
- *Composition styles:* Detailed, Geometric, Minimal, Texture
- *Genre:* Horror, Noir, Romantic, Vintage
- *Mood:* Ethereal, Melancholy, Serene
- *Optical techniques:* Bokeh, Depth-of-Field, HDR, Long Exposure, Macro

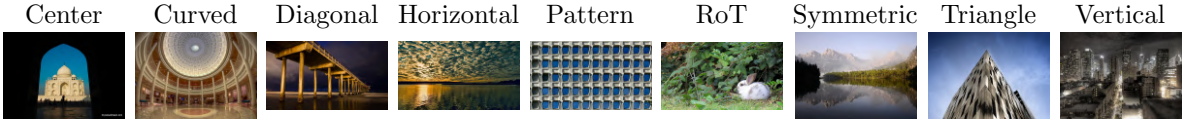
The dataset is split into 64,000 training images and 16,000 testing images. At the time of writing, a total of 63,493 is still downloadable. Thus, the new splits result in 50,868 training images and 12,625 testing images. Images sampled from the dataset are shown in Figure 4.3a.

KU-PCP: The KU-PCP dataset [67] consists of 4,244 outdoor photographs (3,169 for training and 1,075 for testing). It has been annotated by 18 human subject to categorize

Table 4.1: Image attributes available for AADB, AVA, and in the proposed selection.

Attribute Name	AADB	AVA	Proposed	Attribute Name	AADB	AVA	Proposed
Balancing elements	✓			Melancholy			✓
Bright			✓	Minimal			✓
Bokeh			✓	Motion Blur	✓	✓	
Center			✓	Negative Photo		✓	
Color harmony	✓			Noir			✓
Complementary		✓		Object	✓		
Content	✓			Pastel			✓
Curved			✓	Photo Grain		✓	
Depth-of-Field	✓	✓	✓	Pattern/Repetition	✓		✓
Detailed			✓	Romantic			✓
Diagonal			✓	Rule of Thirds	✓	✓	✓
Duotones		✓		Serene			✓
Ethereal			✓	Silhouettes		✓	
Geometric			✓	Soft Focus		✓	
Hazy			✓	Sunny			✓
HDR		✓	✓	Symmetry	✓		✓
Horizontal			✓	Texture			✓
Horror			✓	Triangle			✓
Light	✓			Vanishing Point		✓	
Light on White		✓		Vertical			✓
Long Exposure		✓	✓	Vintage			✓
Macro		✓	✓	Vivid Color	✓		

(a)



(b)

Figure 4.3: Sample images from (a) the style categories of the FlickrStyle database [56] and (b) the geometric composition categories of the KU-PCP database [67].

images into nine not mutually exclusive geometric classes: center, curved, diagonal, horizontal, pattern, Rule of Thirds (RoT), symmetric, triangle, and vertical. Sample images for each category are reported in Figure 4.3b.

Datasets for image aesthetic assessment

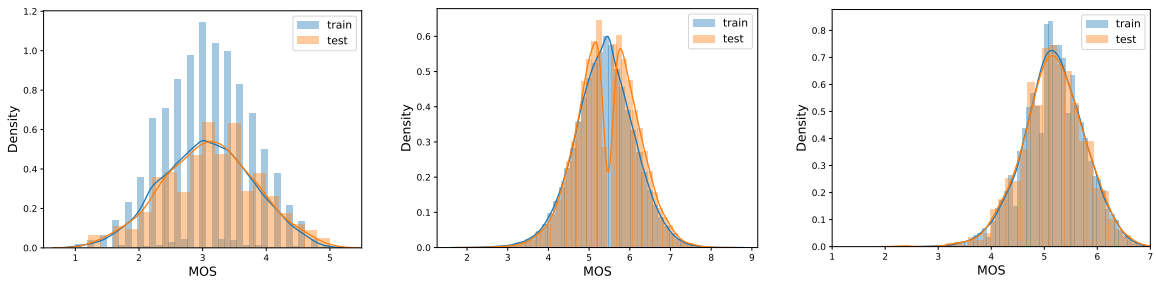
AADB: The Aesthetics and Attributes DataBase (AADB) dataset [63] contains a set of 10,000 images downloaded from the Flickr website.[1] Five Amazon Mechanical Turk (AMT) workers annotate each image with an overall aesthetic score and a set of eleven meaningful attributes. These attributes span traditional photographic principals of color, lighting, focus and composition, and are the following: interesting content, object emphasis, good lighting, color harmony, vivid color, shallow depth of field, motion blur, rule of thirds, balancing element, repetition, and symmetry. For the aesthetic score, AMT workers were allowed to express their judgement on a scale from 1 to 5. For each image, the aesthetic score is the average over all the users judgements. The AADB database was split by its authors into 8,500 images for training, 500 images for validation and 1,000 images for testing. Figure 4.4a shows the distribution of mean ratings for training and test sets.

AVA: The Aesthetics Visual Analysis (AVA) dataset [95] is a large-scale and challenging dataset for image aesthetic assessment. It contains more than 250,000 photos gathered from `www.dpchallenge.com`. Each image provides three types of annotations: aesthetic ratings ranging from 1 to 10 given by about 200 voters; 0, 1 or 2 textual tags chosen from 66 that describe the semantic content of the image; photographic style annotations corresponding to 14 photographic techniques. From the overall set of images, the authors sampled 20,000 for testing (of which only 19,926 are currently available). Following [48], the remaining 235,574 images are further randomly split into training (95%) and validation (5%) sets. Figure 4.4b shows the distribution of mean ratings for training and test sets.

Photo.net: The Photo.net dataset [28] is collected from `www.photo.net`.[2] It contains 20,278 images annotated by at least 10 users to assess the aesthetic quality from one to seven. Of all the images in the dataset, only 16,662 have the distribution of aesthetics ratings and are available. Following [141], from the overall images, 1000 images are used for validation, 1200 images are used for test, and the remaining 14,462 images are used to train. Since the image indexes for each split are not available from [141], the images are randomly divided based on the previous partitioning. To mitigate any bias due to the data division, the partitioning is repeated 10 times and the average performance is reported across the 10 runs.

[1] `http://www.flickr.com`

[2] Available at `http://ritendra.weebly.com/aesthetics-datasets.html`



(a) (b) (c)

Figure 4.4: Distributions of the aesthetic scores on the AADB [63] (a), AVA [95] (b), and Photo.net [28] (c) datasets.

4.4.2 Evaluation Metrics

The proposed method is evaluated with respect to three aesthetic quality tasks (i) aesthetic score regression, (ii) aesthetic quality classification, and (iii) aesthetic score distribution prediction. For the aesthetic score regression task, the mean score of the predicted score distribution is estimated via $\mu = \sum_{i=1}^{N} s_i \times p_i$, with s_i representing the score bucket and p_i that is the estimated probability for the i-th bucket. Following [90, 87, 55], for the aesthetic quality classification, a threshold over the mean score has been applied using the threshold T such that images with predicted score above T are categorized as high quality and vice versa. The evaluation metrics related to the three task are the following:

- Image aesthetic score regression: Results are reported in terms of Spearman's Rank-Order Correlation Coefficient (SROCC), Pearson's Linear Correlation Coefficient (PLCC), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). SROCC measures the monotonic relationship between the ground-truth and the predicted scores, PLCC measures the linear correlation between the actual and the predicted scores. Both SROCC and PLCC range from -1 to 1 and values closer to the extremes indicate better results. RMSE and MAE range from 0 to $+\infty$ and smaller values indicate better results.

- Image aesthetic quality classification: Classification performance is measured in terms of the overall accuracy, defined as $Accuracy = \frac{TP+TN}{P+N}$.

- Image aesthetic score distribution prediction: EMD is used to estimate the closeness of the predicted and ground-truth rating distributions. The EMD is defined in Equation 4.15 with $r = 1$ and lower values of EMD mean better results.

For a detailed description see the section of this thesis A.1 in the appendix.

4.4.3 Training Procedure

The learnable parameters of the proposed model are exclusively θ_s and θ_h, i.e. those of the AttributeNet f_s and those of the HyperNet h. In fact, as previously described in

Section 4.3.1, the θ_b^* parameters belong to a fixed ImageNet-trained Backbone. On the other end, the AestheticNet receives the generated parameters $\hat{\theta}_t$ from the HyperNet.

Similar to [4], a two-step optimization procedure is adopted to introduce attribute-constraint into the HyperNet. First the AttributeNet is trained for aesthetic-related attributes recognition. Then the AttributeNet is froze and the HyperNet is trained for the aesthetic assessment.

Both the AttributeNet and the AestheticNet receive the embedding $\mathbf{e}_b$ produced by the pre-trained Backbone as input. To reduce the training time, as in [48] and [69], the embedding produced by the Backbone for the dataset images are stored instead of calculating them at each training process.

Training for aesthetic-related attributes recognition Multi-Task Learning (MTL) is exploited to train the Multi Layer Perceptron parameters θ_s of the AttributeNet for predicting both image style and image composition. Let $\theta_s = \{\mathbf{W}_s, \mathbf{W}_v, \mathbf{W}_c\}$ represent the weights for the AttributeNet. The bias terms are eliminated for simplicity. Given the dataset $\mathcal{D}_v = \{(\mathbf{X}_v^{(i)}, \mathbf{y}_v^{(i)})\}_{i=1}^M$ for image style recognition and the dataset $\mathcal{D}_c = \{(\mathbf{X}_c^{(i)}, \mathbf{y}_c^{(i)})\}_{i=1}^N$ for image composition recognition, the proposed AttributeNet aims to minimize the combined loss of both tasks:

$$\text{argmin}_{\theta_\mathbf{s}} a_v \sum_{i=1}^{M} \mathcal{L}_v(\mathbf{e}_b^{(i)}, \mathbf{y}_v^{(i)}) + a_c \sum_{j=1}^{N} \mathcal{L}_c(\mathbf{e}_b^{(j)}, \mathbf{y}_c^{(j)}), \tag{4.13}$$

where a_v and a_c control the importance of each task and correspond to 1 and 10, respectively. The embedding $\mathbf{e}_b^{(i)} = b(\mathbf{X}_v^{(i)}; \theta_b^{*(L)})$ $\mathbf{e}_b^{(j)} = b(\mathbf{X}_c^{(j)}; \theta_b^{*(L)})$ are obtained from the Backbone for both the training sets. The dataset $\mathcal{D}_v$ used for style recognition is FlickrStyle. As described in Section **??**, the FlickrStyle categories have been annotated as mutually exclusive therefore cross-entropy is adapted as $\mathcal{L}_v$. The KU-PCP dataset instead is adopted as $\mathcal{D}_c$ training set. It is labeled with nine not mutually exclusive image composition classes. Hence, the binary cross-entropy is used as the $\mathcal{L}_c$ loss function.

The cardinality of the image style recognition dataset is greater than that of the image composition: the FlickrStyle dataset has 50,868 training images, while KU-PCP consists of 3,169 training images. For this reason, during the training phase, the number of images between the two datasets is balanced by performing data augmentation on KU-PCP. Augmentation techniques that do not affect the image composition, i.e., color jittering (random adjustment of brightness, contrast, saturation, hue), random horizontal flipping, random grayscale, and random patch erasing, have been selected.

The size of the embedding vector $\mathbf{e}_s$ is fixed to 512. The learning rate is initially set to $1e^{-4}$ and then dropped by 10 every 20 epochs. A batch size of 32, randomly sampling images from both the KU-PCP and the FlickrStyle, is used. The model is trained for a maximum of 60 epochs using Adam [59] as optimizer monitoring the accuracy over the validation set to select the best model.

Training for aesthetic assessment The second training concerns the θ_h parameters of the HyperNet for generating the parameters $\tau theta_t$ of the AestheticNet, which in turn manages the aesthetic assessment. In this thesis, the aesthetic assessment is formulated as a label distribution prediction problem. More in detail, the proposed network f_t is not trained to predict the Mean Opinion Score (MOS), instead it infers the l_1-normalized score distribution $\hat{\mathbf{q}} = [\hat{q}_{s_1}, \hat{q}_{s_2}, ..., \hat{q}_{s_B}]$. Where s_i is the i-th score bucket, B is the total number of score buckets, and $\hat{q}_{s_i}$ denotes the number of voters that give the discrete score s_i to the image.

Given the dataset $\mathcal{D}_t = \{(\mathbf{X}_t^{(k)}, \mathbf{q}^{(k)})\}_{k=1}^N$, the ground-truth of each image k is represented by a score distribution $\mathbf{q} = [q_{s_1}, q_{s_2}, ..., q_{s_B}]$ defined as above.

The HyperNet is optimized as follows:

$$\text{argmin}_{\theta_h} \sum_{k=1}^{N} \mathcal{L}_{task}(\mathbf{e}_b^{(k)}, \mathbf{q}^{(k)}), \tag{4.14}$$

where $\mathbf{e}_b^{(k)} = f_s(\mathbf{e}_b; \theta_s^{*(M)})$ is the attribute-conditioned embedding obtained from the previously trained AttributeNet for the training image $\mathbf{X}_t^{(k)}$. The loss $\mathcal{L}_{task}$ is the Earth Mover's Distance (EMD). Given the predicted $\hat{\mathbf{q}}$ and the ground-truth $\mathbf{q}$ score distributions, the EMD loss function is defined as follows:

$$EMD(\hat{\mathbf{q}}, \mathbf{q}) = \left(\frac{1}{N} \sum_{k=1}^{N} |CDF_{\hat{\mathbf{q}}}(k) - CDF_{\mathbf{q}}(k)|^r \right)^{\frac{1}{r}}, \tag{4.15}$$

where $CDF_*(k)$ is the cumulative distribution function, r equal to 2 is used to penalize the Euclidean distance between the CDFs.

The AestheticNet consists of $M = 5$ linear layers whose output sizes are 512, 256, 256, 64, respectively. The last linear layer have a number of neurons in output equal to the number of buckets of the score distribution which depends on the training dataset: AADB dataset have a total of 5 buckets ($B = 5$) with $s_1 = 1$ and $s_B = 5$; AVA dataset have a total of 10 buckets ($B = 10$) with $s_1 = 1$ and $s_B = 10$; the Photo.net dataset, $B = 7$, $s_1 = 1$, $s_B = 7$. This training is ran for 40 epochs exploiting the Adam optimizer. The initial learning rate corresponds to $1e^{-5}$, then it is divided by 10 every 20 epochs. The SROCC is tracked over the validation set to select the best model.

4.5 Results

In this section, first the results obtained by the method proposed for aesthetic-related attribute recognition are reported. Then the effectiveness of the proposed method for image aesthetic assessment is measured on the three considered datasets, namely AADB, AVA, and Photo.net. Next, the results are compared with those of many other methods in the state-of-the-art.

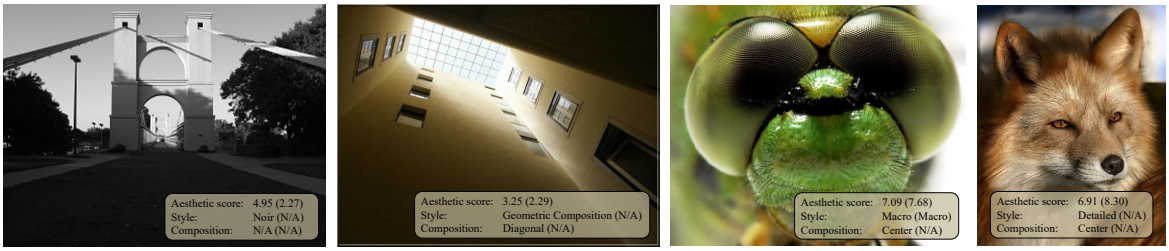


Figure 4.5: Output produced by the proposed method on sample images from the AVA dataset. For each image, the aesthetic score and the attributes predicted by the proposed method are reported (ground-truth is in brackets). "N/A" means that the dataset does not provide any style annotation for the image.

Figure 4.5 shows some sample predictions produced by the proposed method on the AVA test images. For each image, the aesthetic score and style and composition tags predicted by the proposed method as well as the corresponding ground-truth are reported.

4.5.1 Image style and composition recognition

The performances of the proposed method is evaluated over the image style and composition recognition over their relative datasets, respectively the FlickrStyle dataset for the former and the KU-PCP dataset for the latter.

The style labels are learnt and predicted on the 63,493 images available or the FlickrStyle dataset, labelled with 20 different visual styles. Accordingly to the original paper, 20% of the data is used for the test set. Figure 4.6 reports the confusion matrix on the test set. Unfortunately the results are not directly comparable with those the original paper for two main reasons. First, a considerable amount of pictures is missing (about 20% of images with respect to the presented datasets). Second the publishers of the dataset compute the confusion matrix over a random class-balanced subset of the test data (each class has equal prevalence), a subset which is not provided. Notwithstanding, the proposed method shows similar behaviour over the per-class accuracies: the proposed model performs well over styles as Sunny, Noir and Macro while is less effective on attributes like Romantic and Depth of Field. Observing the confusion matrix reported in Fig. 4.6, the proposed model shows understandable confusion over similar styles: Depth of Field vs Bokeh, Horror vs Noir and Pastel vs Vintage.

The performance of the proposed method is evaluated with respect to the image composition recognition on the official test set of KU-PCP dataset. As the authors of the dataset, the performance is quantitatively evaluated measuring the score in terms of accuracy: $Accuracy = \frac{N_c}{N}$ where N and N_c respectively indicate the number of total total and correctly classified photographs. As the publishers of the dataset, an image is considered correctly classified, if it is assigned to at least one of the ground-truth composition classes. The proposed model registers an accuracy of 70.87 which is in line with the performance of the authors training an SVM classifier over the deep features extracted from a CNN

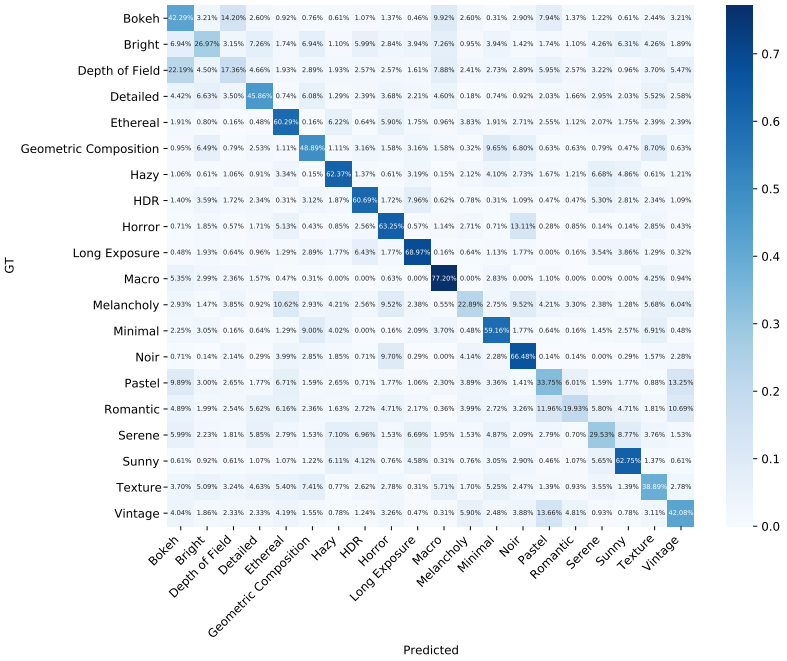


Figure 4.6: Confusion matrix on the Flickr Style categories.

pre-trained on ImageNet, which achieve an accuracy of 70.23%.

4.5.2 Image aesthetic assessment

In this section we summarise the results and compare them with state-of-the-art methods. For each dataset we report the results declared by the authors in the reference papers for the calculated metrics.

As a baseline we considered a dummy automaton that always assigns the average of the training set scores to each test image. More in detail, for each test image, a normal distribution is generated with a mean equal to the average of the training set scores and a standard deviation randomly sampled in the range [-0.5,0.5].

Table 4.2 reports the results on the AADB dataset. From the results it is possible to draw several conclusions. First, the proposed method outperforms all the state-of-the-art methods for the SROCC metric which is the only one to be reported by all methods. Second, the proposed method outperforms Leonardi *et al.* [69] for all metrics. In particular, the accuracy of the proposed method is 2% higher than that of the Leonardi. Finally, the third method, i.e. *RGNet* [78], has a SROCC of 0.03 lower than the proposed method.

In Table 4.3 the results on the AVA dataset are reported. All the state-of-the-art methods, apart from *PI-DCNN* [111] and Pan *et al.* [100], calculated accuracy, thus indicating that they treat aesthetic evaluation as a binary problem. The *Baseline* achieved an accuracy of 71.28% which is 12% lower than the best accuracy corresponding to 83.59% for *RGNet*. Interestingly, no method achieves the best performance for all metrics. This

Table 4.2: Comparison of the proposed method with state-of-the-art methods on the AADB dataset. The "–" means that the result is not available.

Network architecture	Accuracy (%) ↑	SROCC ↑	PLCC ↑	MAE ↓	RMSE ↓	EMD ↓
Baseline	61.58	-0.0744	-0.0543	0.1449	0.1799	0.1407
Reg-Net (AlexNet) [63]	–	0.6782	–	–	–	–
Malu *et al.* (ResNet-50) [91]	–	0.6890	–	–	–	–
PI-DCNN (ResNet-50) [111]	–	0.7051	–	–	–	–
Chen *et al.* (ResNet-50) [20]	–	0.7080	–	–	–	–
Pan *et al.* (ResNet-50) [100]	–	0.7041	–	–	–	–
Reddy *et al.* (EfficientNet-B4) [106]	–	0.7059	–	–	–	–
RGNet (DenseNet-121) [78]	–	0.7104	–	–	–	–
Leonardi *et al.* (EfficientNet-B4) [69]	79.51	0.7454	0.7479	0.1062	0.1351	–
Proposed	**81.64**	**0.7567**	**0.7616**	**0.0832**	**0.1059**	**0.0951**

makes it difficult to understand which method is the best. *RGNet* obtains the best accuracy, while MLSP [48] shows the highest regression metrics against mid-ranking accuracy (81.68%). The proposed method ranks second for the regression metrics, first for the EMD metric, while it achieves an accuracy 3% lower than the 83.59% by *RGNet*.

Figure 4.7 shows ten samples from the AVA test set predicted by the proposed method as having high aesthetic quality (the top five images) and low aesthetic quality (the bottom five images), respectively. Plots of the ground-truth and predicted distributions are also shown. As it is possible to see, the model can achieve a high degree of accuracy, with an estimate of the score distribution almost perfect in some cases.

In Figure 4.8 there are two examples of failure of the proposed method on the AVA test set images. The method behaves badly on images with not very Gaussian score distributions. This might depend on the fact that 99.77% of the images in the dataset instead follows a Gaussian distribution [95].

Finally, Table 4.4 shows the comparison on the Photo.net dataset. The results for state-of-the-art methods are reported by [141]. The performance is not directly comparable. The evaluation protocols adopted for the methods are different (see the column "Evaluation protocol" for details about the adopted protocols). The performance of the proposed method is comparable to that reported for *GPF-CNN* that is the method achieving results similar to ours.

Several things can be deduced from the results. First, the *Baseline* accuracy is very high (66.58%) compared with the average obtained by the methods on this dataset. It exceeds that of three methods, i.e. *GIST_SVM* [93], *FV_SIFT_SVM* [93], and *MTCNN* [55]. Second, the proposed method records a significant improvement on all metrics apart from accuracy compared to *GPF-CNN*. In particular, the SROCC of 0.5650 is 0.04 higher than that of *GPF-CNN*, the MAE is 0.05 lower. Third, the small standard deviation indicates that the proposed method is able to generalize well.

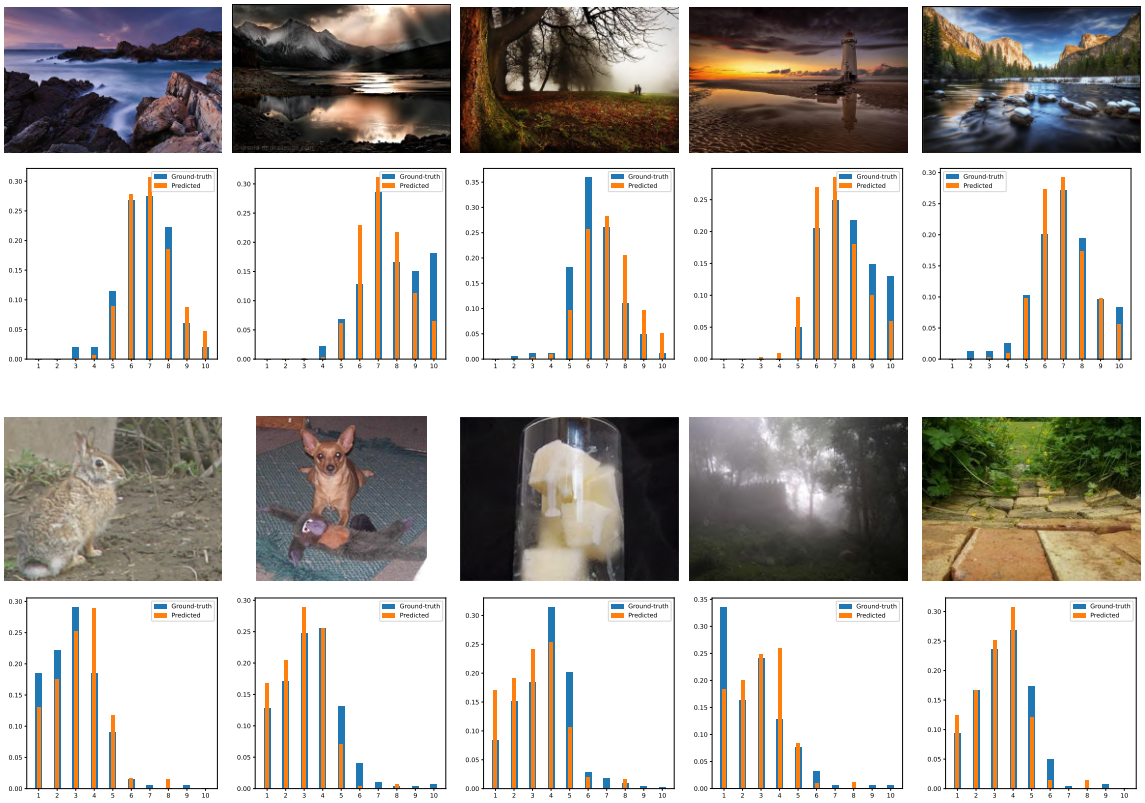


Figure 4.7: Sample predictions by the proposed method on AVA test images. Top 2 rows: predicted images with high aesthetic quality, coupled with plots of their ground-truth and predicted score distributions. Bottom 2 rows: predicted images with low aesthetic quality, coupled with plots of their ground-truth and predicted score distributions.

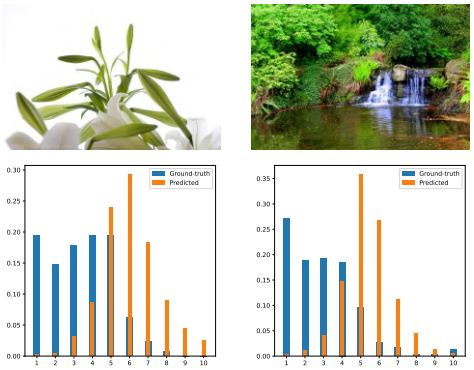


Figure 4.8: Failure cases of the proposed method on the AVA test set images.

Table 4.3: Comparison of the proposed method with state-of-the-art methods on the AVA dataset. In each column, the best and second-best results are marked in **boldface** and underlined, respectively. The "–" means that the result is not available.

Network architecture	Accuracy (%) ↑	SROCC ↑	PLCC ↑	MAE ↓	RMSE ↓	EMD ↓
Baseline	71.28	-0.0003	-0.0021	0.6230	0.7550	0.0743
RAPID (AlexNet) [82]	74.20	–	–	–	–	–
DMA-Net (AlexNet) [83]	75.42	–	–	–	–	–
MNA-CNN (VGG16) [90]	76.10	–	–	–	–	–
Reg-Net (AlexNet) [63]	77.33	0.5581	–	–	–	–
MTCNN (VGG16) [55]	78.56	–	–	–	–	–
Multimodal DBM Model (VGG16) [143]	78.88	–	–	–	–	–
NIMA (VGG16) [118]	80.60	0.5920	0.6100	–	–	0.0520
GPF-CNN (VGG16) [141]	80.70	0.6762	0.6868	0.4144	0.5347	0.0460
NIMA (InceptionNet) [118]	81.51	0.6120	0.6360	–	–	0.0500
MLSP (InceptionNet) [48]	81.68	**0.7524**	**0.7545**	**0.3831**	**0.4943**	–
GPF-CNN (InceptionNet) [141]	81.81	0.6900	0.7042	0.4072	0.5246	0.0450
MULTIGAP (InceptionNet) [46]	82.27	–	–	–	–	–
A-Lamp (VGG16) [87]	82.50	–	–	–	–	–
AFDC+SPP [19]	83.24	0.6489	0.6711	–	–	0.0447
PI-DCNN (ResNet-50) [111]	–	0.6578	–	–	–	–
Pan *et al.* (ResNet-50) [100]	–	0.7041	–	–	–	–
RGNet (DenseNet-121) [78]	**83.59**	–	–	–	–	–
Proposed	80.75	0.7318	0.7329	0.4011	0.5128	**0.0439**

Table 4.4: Comparison of the proposed method with state-of-the-art methods on the Photo.net dataset. In each column, the best and second-best results are marked in **boldface** and underlined, respectively. The "–" means that the result is not available.

Network architecture	Accuracy (%) ↑	SROCC ↑	PLCC ↑	MAE ↓	RMSE ↓	EMD ↓	Train-val-test remarks
Baseline	66.58 ± 0.30	0.0060 ± 0.02629	0.0053 ± 0.0285	0.4481 ± 0.0023	0.5652 ± 0.0023	0.0789 ± 0.0004	15K train, 1000 val, 1200 test*
GIST_SVM [93]	59.90	–	–	–	–	–	5-fold cross-validation
FV_SIFT_SVM [93]	60.80	–	–	–	–	–	5-fold cross-validation
MTCNN (VGG16) [55]	65.20	–	–	–	–	–	about 15K train, 3000 test
GPF-CNN (VGG16) [141]	**75.60**	0.5217	0.5464	0.4242	0.5211	0.0700	15K train, 1000 val, 1200 test
Proposed	70.05±0.89	**0.5650±0.0153**	**0.5698±0.0141**	**0.3714±0.0065**	**0.4700±0.0071**	**0.0689±0.0009**	15K train, 1000 val, 1200 test*

* Average and standard deviation on the 10 iterations of train-val-test splits.

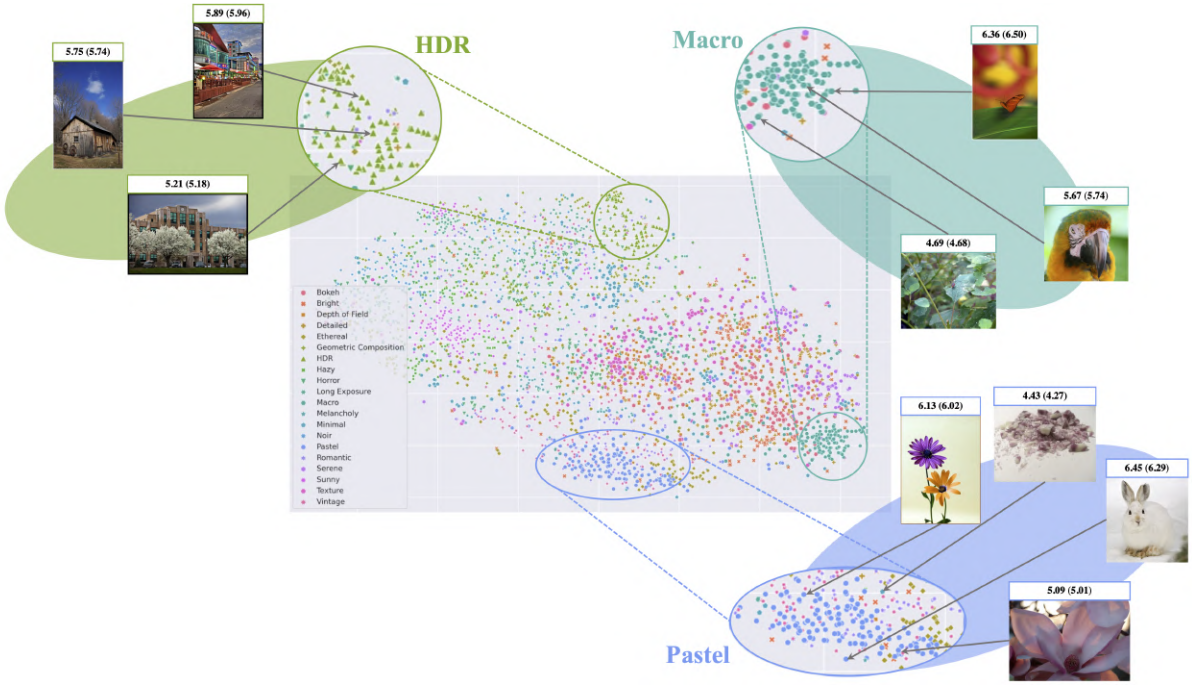


Figure 4.9: The predicted weights for several images of the AVA test set are plotted in the 2D space after the t-SNE transformation. This figure shows the weights extracted from the last layer of the target network, the weights of the other layers also show a similar distribution. For each of the depicted images, the predicted aesthetic score (ground-truth is in brackets) is reported.

4.5.3 Visualization of predicted weights

To assess the influence of the style of the images over the aesthetics prediction, the weights generated by the HyperNet are extracted from several images of the AVA test set. All of the 19,926 images have been labelled with style attributes from the FlickrStyle dataset with a model trained as stated before: first the AttributeNet is trained for aesthetic attributes recognition on both the FlickrStyle and KU-PCP datasets. Then the HyperNet is trained for the aesthetic assessment over the AVA dataset. Then 200 images have been randomly selected for each of the 20 style attributes and store the weights of the AestheticNet's last linear layer predicted by the HyperNet. Then the dimension of the predicted weights has been reduced with t-distributed stochastic neighbor embedding (t-SNE) [122] and plotted them in a 2D space for visualization. In Figure 4.9, are shown the transformed weights extracted from the last layer of the HyperNet, weights from other layers also exhibit similar distribution.

From Figure 4.9, several interesting behaviours can be observed. First, for different images the generated weights vary. This indicates that the proposed method adopt distinct weights for evaluating image aesthetics in a self-adaptive manner. Whereas for traditional automatic aesthetics assessment models the weights are fixed for all input images.

Secondly, images belonging to the same image style generate weights that are close to each other. This verifies that the training of the AttributeNet is effective.

Furthermore, it is noticeable how the information encoded in the embedding produced by the AttributeNet is successfully propagated through the whole model. As can be seen from Figure 4.9, even though their image aesthetics vary, images sharing the same style (e.g. HDR, paste, macro) generate similar weights for image aesthetics assessment.

This encoding of the image style in the proposed model is believed to makes it self-adaptive in the direction of understanding the image aesthetics through the attributes that describe its artistic aspect and the photography rules applied when shooting. Thus, the proposed model is believed to be more flexible and exhaustive in the evaluation of image aesthetics.

Chapter 5

Image Memorability using Diverse Visual Features and Soft Attention

5.1 Introduction

A remarkable feature of human cognition is the ability to remember different images that have been seen only once [64]. Furthermore, different people tend to remember or forget same pictures. This result suggests that people encode and discard very similar types of information. Precisely, images that are usually forgotten seem to lack distinctiveness and a fine-grained representation in human memory [64]. Taking into account the aforementioned considerations, it seems that memorable images have some kind of intrinsic visual features, making them easier to remember. Indeed, past studies have shown that memorability is a measurable stationary property of an image shared across different viewers [52] and that it is possible to determine a compact set of attributes characterizing the memorability of any individual image [51]. These results led researchers wondering how to predict accurately which images will be remembered and which will be not, resulting in the first large scale visual memorability estimation with near-human performance [58].

Nowadays, we are continuously being exposed to photographs when browsing the Internet or leafing through a magazine. Exploiting memorable pictures can have a huge impact in many applications, also thanks to the relationship between emotions and memorability [13]. Just to give some examples, estimating the memorability can help to automatically select the images that can have a key role in optimizing the conversion rate for media advertisement and online shopping, or in improving the communication of a specific concept. More recently, researchers started to show interest in how to make an image more memorable, by exploiting deep architectures for generating memorable pictures by exploiting style-transfer techniques [112].

In this thesis, we proposed a novel approach to compute memorability that exploits the combination of feature computed by different Convolutional Neural Networks (CNNs) and an attention map extracted from a caption generation model with visual attention. In details, the main contributions of the approach are:

- To achieve comparable or better results with respect to state-of-the-art approaches, respectively in terms of Spearman's rank correlation and Mean Squared Error (MSE);
- To reduce the amount of parameters with respect to the best performing technique in terms of Spearman's rank correlation.

5.2 Related work

In the first works on image memorability, Isola *et al.* [51, 52] showed the ability of our mind to remember certain images better than others and also that memorability is a stable property across different viewers. They introduced a database for which they collected the probability that each image will be remembered after a single view as well as image attribute annotations (such as spatial layout, content and aesthetic properties) in order to:

- Understand which features are highly informative about memorability;
- Demonstrate that memorability is not influenced by content frequency or familiarity, namely the presence of particular objects, scene categories, relatives or famous monuments. However, some contents like faces are memorable, while vistas and peaceful settings are not;
- Prove that memorability is not correlated with aesthetics, interestingness, and simple image features.

Furthermore, they developed a method to predict the memorability of an image involving the use of Support Vector Regressor machines on the combination of global image features – GIST [98], SIFT [66], HOG [27], SSIM [110], and color histogram. Following the intuition of Isola *et al.* [51] that memorability and visual attention are correlated, Mancas and Le Meur [92] demonstrated that attention-related features can effectively replace some of the low-level features used by Isola *et al.* [52] and thus reducing the dimensionality of the feature set. Afterwards, Bylinskii *et al.* [12] proved that the interplay between intrinsic image properties (the fact that some scene categories are more memorable than others) and extrinsic factors, such as image context and observer behavior, are necessary to build an improved image memorability model. The effectiveness of the proposed solution has been assessed on FIne-GRained Image Memorability (FIGRIM) dataset that is composed by more than 9K images.

Khosla *et al.* [58] released LaMem, the first large scale dataset for image memorability containing 60K images. Alongside the dataset, they proposed MemNet, a CNN for memorability score estimation. The model is based on the fine-tuning of Hybrid-CNN [142], a CNN trained using 3.5 million images from 1,183 categories, obtained by merging the scene categories from Places database [142] and the object categories from ImageNet [108]. They achieved near human consistency rank correlation (0.68) for memorability. Fajtl *et al.* [33] proposed AMNet, a model consisting of a ResNet50 [44] pre-trained on

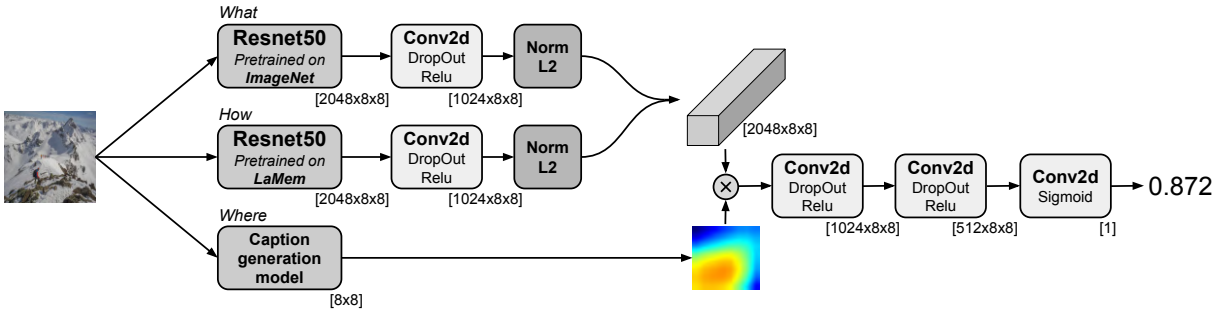


Figure 5.1: Overview of the proposed model for image memorability estimation. The attention map produced by the caption generation model is combined channel-wise with the feature volume.

ImageNet, a soft attention mechanism, and a Long Short-Term Memory [42] for memorability score regression. The AMNet model achieved a performance of 0.677 in terms of Spearman's rank correlation on LaMem dataset. Recently, Squalli-Houssaini *et al.* [116] approached the task of image memorability estimation as a classification problem instead of a regression one. They developed a model combining features extracted from both a VGG16 [114] pre-trained on ImageNet and an image captioning system [60] and outperformed both state-of-the-art and human consistency correlation (0.72) on LaMem dataset.

5.3 Proposed method

Image memorability is influenced by some intrinsic image properties, namely *what* kind of objects and scenes are present and *what* are their characteristics, but also by extrinsic factors such as the image locations *where* humans focus their attention. Our approach tries to model memorability according to the aforementioned aspects by using a CNN for encoding intrinsic characteristics of objects, and a soft attention mechanism for estimating attention maps that highlight salient regions. Furthermore, in the proposed model, a CNN pre-trained on image memorability is include for mapping *how* features encode memorability.

5.3.1 Architecture

The proposed model, depicted in Figure 5.1, estimates a memorability score given as input an RGB image of size 256×256 pixels. It consists of two CNNs trained on two different tasks, and a soft attention mechanism based on a system originally designed for caption generation [132]. The aforementioned blocks (i.e. soft attention and memorability) are followed by two convolution layers preceding the last regressor module, which estimates the memorability score.

Feature extraction The two considered CNNs are two ResNet50 architectures pre-trained respectively for: image memorability estimation on LaMem dataset [58] and object recognition on Imagenet [108] dataset. These two CNNs are considered to provide the model prior information over the memorability of the image as well as knowledge of the image context. Both the architectures are truncated before their last average pooling layer in order to obtain two feature maps of size $2048 \times 8 \times 8$. These feature maps are first passed through a convolution layer which halves their channel dimension, then they are L2-normalized by dividing the feature map by its L2-norm, and finally stacked together obtaining a new feature map having a dimension equal to $2048 \times 8 \times 8$.

Soft attention mechanism To focus the model attention on salient regions that are highly informative for memorability estimation, in the proposed model is included a state-of-the-art captioning generation approach [132] for extracting attention maps. This captioning generation model is trained on the MS COCO dataset [77] and produces at most 50 attention maps with spatial size 8×8 pixels, each one focusing on a particular detail of the image. These maps are exploited by averaging them in order to get a single and global attention map.

Memorability estimation The feature map extracted from the two CNNs is weighted with the attention map generated from the captioning model replicated channel-wise. Finally, the resulted weighted feature map is given as input to a three-layer CNN to predict the memorability score.

5.3.2 Training procedure

In order to improve the generalization of the model and minimize the risk of overfitting, data augmentation techniques are used during the training phase. Specifically, random scaling in the range $[0.8, 1.2]$ is first applied to the image, which then is randomly flipped along the vertical axis. Subsequently, random crop (0.8 to 1.0) of the image is applied before sub-sampling it to a size of 256×256 pixels. Finally, the image is normalized by subtracting and dividing each image by the mean and standard deviation estimated on the ImageNet training set [108] in order to limit the variability of the input range.

The training procedure consists of two phases. First one ResNet50 is trained from scratch on LaMem [58] dataset for image memorability. Then the whole model is fine-tuned on the same dataset freezing the weights of the two ResNet50 and the weights of the caption generation model with visual attention [132]. Both of the training processes are trained to minimize the mean squared error between the ground-truth and the predicted image memorability scores. For the first stage, the model is trained for 150 epochs due to a larger number of parameters to learn, with a batch size of 10 images. For the second phase, the model is trained for only 50 epochs with a bigger batch size of 16 images.

During both the training processes, the technique of early stopping is being used analyzing the Spearman's rank correlation (see in the section of this thesis A.1 for the

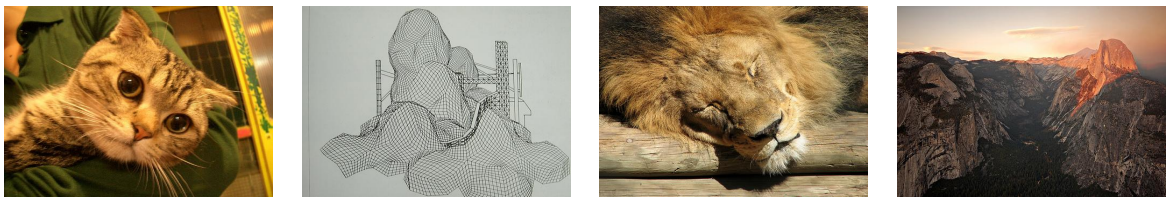


Figure 5.2: Sample images from the LaMem dataset [58].

definition) on the validation set. For both stages the ADAM optimizer [59] is used with starting learning rates respectively of 5×10^{-7} and 5×10^{-5} for the first and the second stage. Both the learning rates are decreased every epoch as follows:

$$LR(epoch) = \left[1 - \left(\frac{epoch}{total\ epochs}\right)^{0.9}\right] * LR_0, \tag{5.1}$$

where *epoch* is the 0-based index of the actual epoch, LR_0 is the initial learning rate, and *total epochs* is the total number of epochs for the training process.

5.4 Experiments

In the following sections, the dataset and metrics adopted for evaluating the proposed method are described. Experimental results are then reported. The proposed approach have been developed using the PyTorch framework [102], on a NVIDIA GTX 1070 GPU.

5.4.1 Dataset

The proposed model is evaluated on the LaMem dataset [58], a collection of 58,741 images annotated with a memorability score. The images were sampled from different existing datasets and cover various indoor and outdoor scenes. Figure 5.2 shows some samples from the dataset. The provided memorability score were collected on Amazon Mechanical Turk using an improved version of the memorability game introduced in [52]. The data are divided into five random training, validation and test set splits. Each of these splits has respectively 45k images as training set, 3741 as validation and 10k as test set. For each split, training and validation sets are labeled from the same group of people while the test is labeled from a different group.

5.4.2 Evaluation Metrics

Following [58], the performance of the proposed method are evaluated using the SROCC, [103] and the MSE previously introduced in the section .

Table 5.1: Results of the ablation study on the LaMem dataset reported in terms of Spearman's rank correlation (SROCC) and Mean Squared Error (MSE).

Method	SROCC ↑	MSE ↓
ResNet50-LaMem	0.680	0.0083
ResNet50-LaMem + ResNet50-ImageNet	0.686	0.0080
Whole model	0.687	0.0079

Table 5.2: Comparison with state-of-the-art methods in terms of Spearman's rank correlation and MSE on the LaMem dataset. For each model the number of its parameters (in millions) is also reported.

Method	SROCC ↑	MSE ↓	# parameters
AMNet [33]	0.677	0.0082	**39M**
MemNet [58]	0.640	N/A	62M
Squalli *et al.* [116]	**0.720**	0.0092*	280M
Proposed model	0.687	**0.0079**	130M

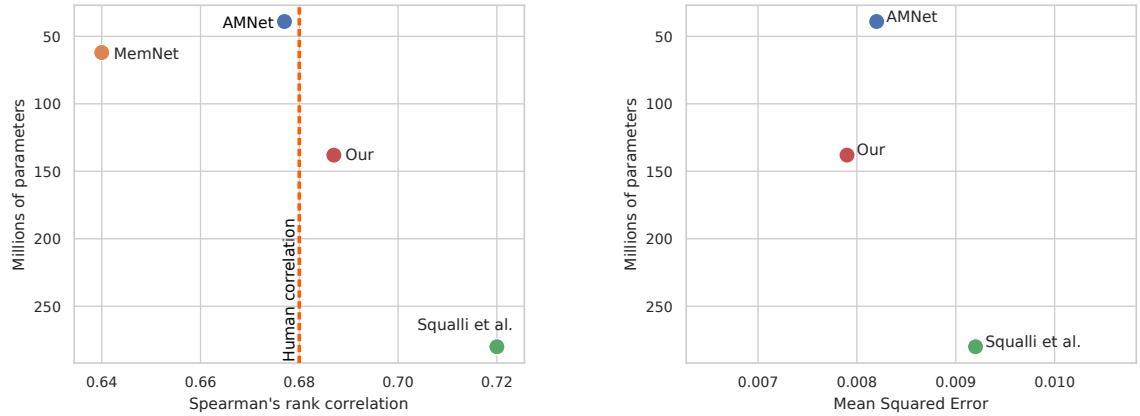
*Estimated by the authors.

5.5 Results

In this section the performance of the proposed model is reported by averaging both the Spearman's rank correlation and MSE over the five splits of LaMem dataset. The proposed model reaches an average rank correlation of 0.687 and a MSE of 0.0079 over the five splits of LaMem.

Table 5.1, reports the results of an ablation study investigating how each module of the proposed model affects the overall performance. In particular, a single ResNet50 [44] trained on the task of image memorability achieves a Spearman's rank correlation of 0.680 and a MSE of 0.0083. The model involving the combination of the feature maps extracted from the two ResNet50 without the use of the attention map increases the correlation by 0.006 and lowers the MSE by 0.0003. Finally, it can be seen that the whole model, i.e. the addition of the soft attention mechanism, increases performance by 0.001 for the Spearman's rank correlation and decreases the MSE by 0.0001.

Table 5.2 compares the proposed method with the state-of-the-art on LaMem [58] dataset. The performance provided are reported in terms of correlation and MSE as the average results over the five dataset splits. From the results reported in Table 5.2 it can be observed that in terms of Spearman's rank correlation, the proposed model performs slightly worse with respect to the best state-of-the-art model [116]. Given that Squalli *et al.* [116] do not provide the MSE, their solution have been implemented and it obtained an error of 0.00923. Based on this result, the proposed approach reduces the MSE by 0.0013 using a number of parameters equal to less than half of those used by [116]. Furthermore have been conducted an analysis of the efficiency of proposed solution respect to previous methods. To this end, Figure 5.3a compares the Spearman's rank correlation

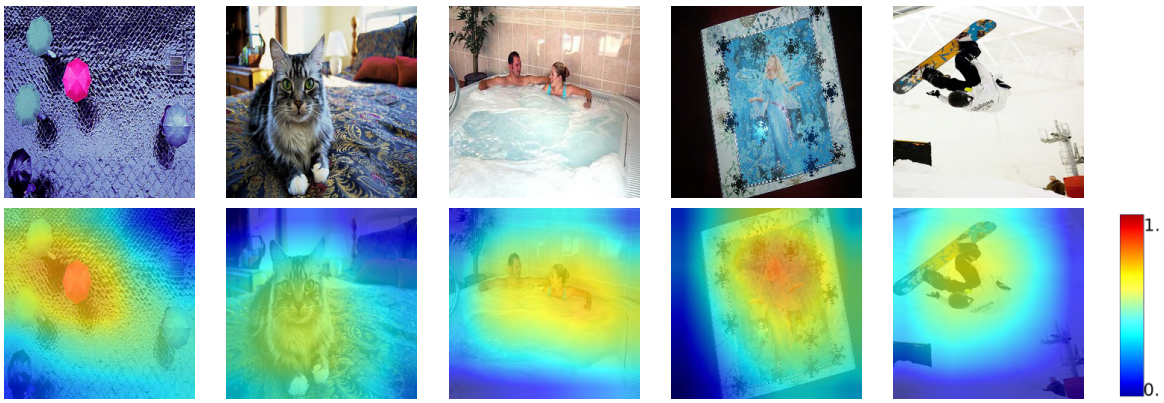


(a) (b)

Figure 5.3: Spearman's rank correlation vs. model parameters (the dashed line depicts the human consistency rank correlation [57]) (a). MSE vs. model parameters (b).

and the number of parameters, while Figure 5.3b plots the MSE and the number of parameters. Among the methods that outperform the human consistency correlation (0.68), the proposed model achieves lower performance by using a reduced amount of parameters. Instead in terms of MSE, the proposed method is the solution exploiting more efficiently its parameters by obtaining the smallest MSE with the fewest parameters.

Figure 5.4 shows samples from LaMem dataset with memorability scores estimated by the proposed solution as well as ground-truth memorability scores. Furthermore, the corresponding attention maps are provided for each image to highlight how these maps in most cases focus on the relevant subjects in the scenes.



a) 0.805 (0.804) b) 0.739 (0.744) c) 0.890 (0.892) d) 0.791 (0.787) e) 0.782 (0.778)

f) 0.782 (0.784) g) 0.839 (0.838) h) 0.824 (0.829) k) 0.615 (0.614) l) 0.692 (0.694)

Figure 5.4: Sample images from LaMem dataset with estimated and ground-truth (in brackets) memorability scores. Below each image its depicted the related visual attention map produced by the caption generation model.

Chapter 6

A general purpose method for image collection summarization and exploration

6.1 Introduction

Nowadays nearly everybody takes photos, for any reason and at any time. With the growth of the smartphone industry, anyone has a high-quality camera in their pocket, and tacking a picture it is just one click away. Images are a way of communication, and most of the online social networks are now dominated by media content. While the prices of the storage are decreasing, the number of photos stored is increasing, leading to collections of images which sizes begin to be a barrier for relieving the captured moments and exploring them, we are submerged by images: in 2015, 300 million photos were published to Facebook [135] daily on average while in 2016, an average of 80 million photos were shared every day on Instagram [50].

The problem of exploring large collections of images is also relevant in the development of deep learning algorithms for signal or image recognition. Deep Neural Architectures require large amount of data for the training process, for instance Imagenet [29] or Microsoft Common Objects in Context [77] contain million of images. Surfing those kinds of image collections, just to simply have a clue on which data we are working on, is basically impossible. It is therefore clear that methods for the automatic selection of the most important and diverse images from a collection of pictures are needed.

Here we propose a flexible framework that can be used both to explore large scale image datasets and summarize photo albums. The proposed approach aims at summarizing image collection by maximizing the diversity between images selected and their quality and/or aesthetic rate.

In figure 6.1 is reported as example the summarization of the pictures included in the Automatic Triage for a Photo Series dataset [18]. In the figure is highlighted how the images are first grouped according to the semantic content, and later how the represen-

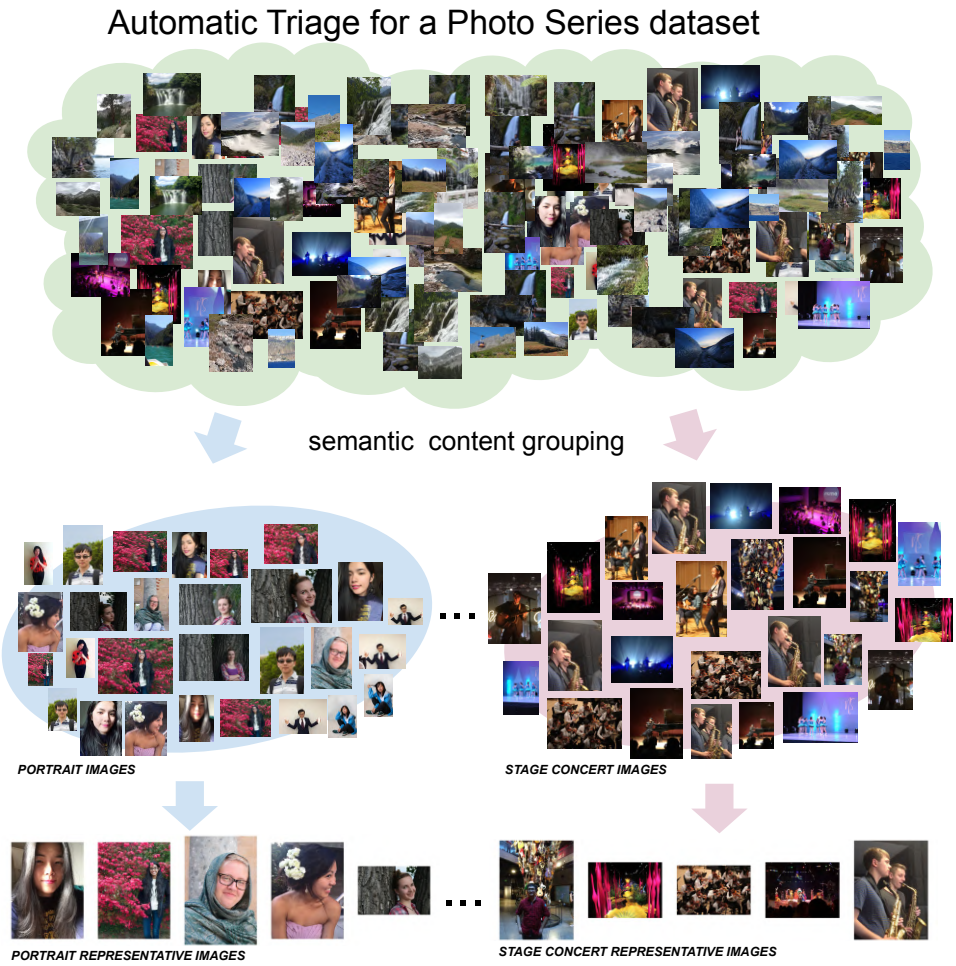


Figure 6.1: Summarizzation example of the Automatic Triage for a Photo Series dataset produced by the proposed method. In the first phase, images are divided into groups with homogeneous semantic content. Subsequently, the most representative pictures are extracted from each of the groups.

tative images are selected from each group. The final summary is therefore given by all the representative images of all the semantic content groups.

To summarize, the contribution of this thesis are the following.

- To present a framework capable of summarizing large collection of images taking into account the diversity and the quality of the resulting set.
- To evaluate the proposed method with respect to various baseline methods.
- To demonstrate the effectiveness of the proposed framework over a user study.

6.2 Related work

Automatic photo selection or album summarization have gained much attention in recent years, and has been studied by several researchers. A simple yet effective way to automatically summarize large photo collection is to divide images into subs groups, for example using algorithms like mean shift [21] or K-means [88], and then select one picture for each group. Following this direction Li *et al.* [71] proposed an automatic organization framework for photo collections based on image content, exploiting a hierarchical clustering technique.

A possible way to improve solutions based on clustering is to model the concept of diversity, for example by adopting solutions that aim to maximise the contrast of the selected subset, like the one proposed by Campadelli *et al.* in [14], where they solve the problem of selecting high-contrast set as combinatorial optimization problem on graphs.

Yeh *et al.* [134] proposed a personalized ranking system for amateur photographs based on aesthetics rules (e.g. Rule of Thirds, Simplicity, Clarity, Color Harmonization) to automatically ranking photographs taking into account individual preferences.

Li *et al.* [72] propose in their manuscript an automatic selection system based on the aesthetic quality of consumer photos, focused solely on photos with faces.

The problem of personal photo collections summarization has been formalized for the first time by Sinha *et al.* [115]. In particular, they state that the problem of photo collections summarization can be formalized over three salient properties, respectively the *quality*, *diversity* and *coverage* that an informative summary should satisfy. In their manuscript, they define the *quality* of a photo summary as the interestingness or attractiveness of the photos present in it, the *diversity* as a measure of its non-redundancy, and the *coverage* as a property that reflects how many important concepts are present in the photolog are also represented in the summary. They therefore, propose a summarization framework that optimizes these properties to generate an informative overview.

Other works [126, 127, 17] focused on photos albums and in particular considered the various possible events that characterize those albums. In [17] Ceroni *et al.* propose an expectation-oriented photo selection method, which combines a variety of image-related factors such as image quality, presence of faces, concept features, and collection based features such as album size. Wang *et al.* [126] claim that the selection process is influenced by the event type, therefore they propose a selection method that takes into consideration the event type, and vary its decision.

6.3 Proposed method

Given a collection of images $\mathcal{H}$, the proposed method selects the most representative images on the basis of several criteria: diversity, quality and aesthetics. The pipeline is divided in three main components, respectively *group selection*, *clustering of pictures within a group* and *best picture selection*.

The first module (*group selection*) aims to group images into L groups $\{\mathcal{G}_1, \cdots, \mathcal{G}_L\}$

based on the semantic content. The choice of the semantic classes is arbitrary and should depend on the kind of task. For example, if we are exploring a dataset, a possible set of semantic classes is the one proposed by Imagenet [29]. Otherwise, if we want to summarize a photo album, scene categories as the one proposed in the Camera Scene Detection Dataset (CamSDD)[105] are more suitable since they cover the most common scenes that can be found in a photo album.

Subsequently, each group of images is processed by the *clustering of pictures within a group* and *best picture selection* modules. In particular, in the first stage, the images of a group $\mathcal{G}_i$ are divided into k sets of pictures $\{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_k\}$ on basis of visual similarity. Afterwards, the best representative picture of the group $\mathcal{C}_j$ is selected on the basis of three perceptual properties: image quality, image aesthetics and *object emphasis* (whether the image emphasizes foreground objects). Therefore each group $\mathcal{G}_i$ is represented by the most k emblematic and diverse images, an the whole set $\mathcal{H}$ is made up of $L \times k$ pictures.

Figure 6.2 shows the schematic overview of the presented pipeline for selecting the best and diverse k photos representing a set of images.

Group selection

In the first step, given a collection of images $\mathcal{H}$, the images are grouped on the basis of the semantic content in L groups $\{\mathcal{G}_1, \cdots, \mathcal{G}_L\}$. Some examples of semantic classes considered are: animals, mountain, architecture, etc. Each group is therefore composed of n pictures $\mathcal{G}_l = \{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_n\}$ such that $\bigcup_{i=1}^{L} \mathcal{G}_i = \mathcal{H}$ and $\mathcal{G}_i \cap \mathcal{C}_j = \emptyset,\ \forall i, j \in \{1, ..., L\} | i \neq j$.

Clustering of pictures within a group

In the second step, given a group of images $\mathcal{G}_l$, the proposed method clusters the pictures using the K-Means algorithm [88] into k clusters $\{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_k\}$ by exploiting the features extracted from the last convolutional layer, after the global average pool, of a ResNeXt-101 32x8d [131]. To better encode the image content, the ResNeXt-101 is pre-trained in weakly-supervised fashion [89] on 940 million public images with 1.5K hashtags [29], followed by fine-tuning on ImageNet1K dataset. This phase ensures the diversity selection creating k clusters of visually similar images such that $\bigcup_{i=1}^{k} \mathcal{C}_i = \mathcal{G}_l$ and $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset,\ \forall i, j \in \{1, ..., k\} | i \neq j$.

Best picture selection

Finally, for each cluster $\mathcal{C}_i$ of a given group $\mathcal{G}_l$, the best image is selected with respect to three perceptual properties, quality, aesthetics and emphasis. Given an image $\mathbf{I}$, let indicate $a(\mathbf{I})$ its perceived aesthetics, $q(\mathbf{I})$ its visual quality and $u(\mathbf{I})$ its grade of *object emphasis*. Define $b(\mathcal{C})$ as the function that, given a group of images $\mathcal{C}$, it returns the best image in $\mathcal{C}$:

$$b(\mathcal{C}) = \arg\max_{\mathbf{I} \in \mathcal{C}} \frac{a(\mathbf{I}) + q(\mathbf{I}) + u(\mathbf{I})}{3}$$

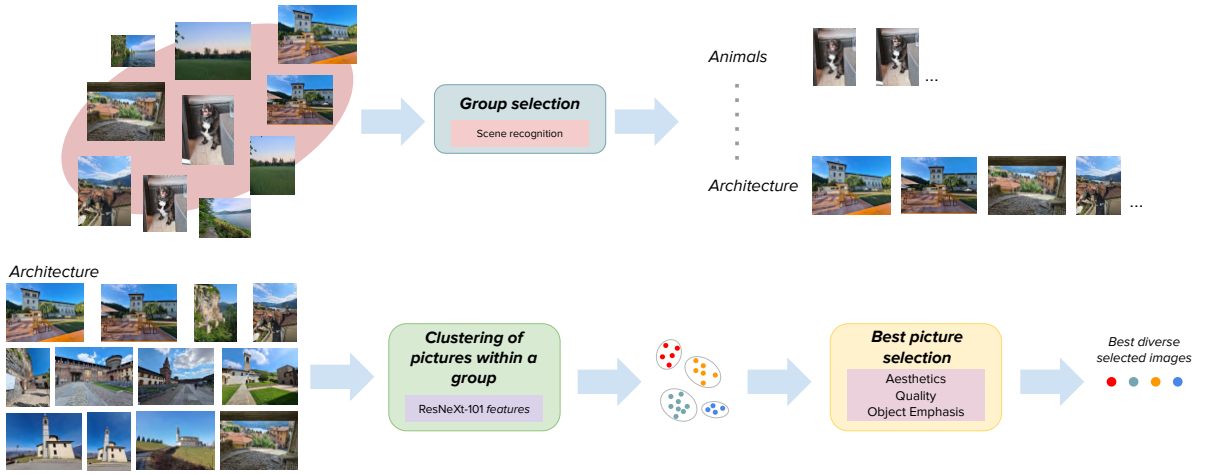


Figure 6.2: Overall pipeline of the proposed framework. Given a collection of photos, first images are divided into homogeneous semantic content groups. Then for each group, the photos are grouped in k clusters adopting the K-means algorithm over features extracted from a ResNeXt-101. Subsequently, for each cluster the best image is elected according to the image aesthetics, quality and the emphasis to the subject of the photo (*object emphasis*).

The best subset $\mathcal{B}$ of images representing $\mathcal{G}$ can be defined as:

$$\mathcal{B} = \{b(\mathcal{C}_i) | 1 \leq i \leq k\}$$

6.4 Experiments

In this section, first the considered datasets are described, then the evaluation metrics used to estimate the performance and the training procedure of the proposed method are described.

6.4.1 Datasets

Choosing the right database is crucial when developing any image summarization solutions. Unedited and complete photo collections are hard to harvest online. Most online photo sharing sites (e.g Flickr), typically contains images that users have chosen to share, therefore they have already been selected from a collection of photos. Even though there already exist several publicly available datasets for the training and testing of image summarization methods, they all have significant limitations.

Some of the most recent datasets are derived from the Yahoo Flickr Creative Commons 100M Dataset (YFCC100M)[121]. The YFCC100M contains 100 million images and videos gathered from Flickr. Each image in the dataset is accompanied by useful

metadata which comprises user ID, tags, and the timestamp of the image. One of the most popular datasets gathered from the YFCC100M is the CUration of Flickr Events Dataset (CUFED) [126]. It is an event curation dataset composed of 1883 albums. It covers a total of 23 most common events of daily life, ranging from Wedding to Nature Trip, and each album contains between 30 and 100 images. Although it includes a considerable amount of images, the CUFED dataset is strictly related to events, and since the photos are collected from Flickr, is therefore uncommon to find several repeated images. Another similar dataset, based on the YFCC100M, has been proposed in [113].

Given the difficulty of finding a dataset composed of images collections, with several repeat pictures, the Automatic Triage for a Photo Series dataset [18] is exploited grouping the photo series according to their camera scene.

Automatic Triage for a Photo Series

The Automatic Triage for a Photo Series dataset [18] consists of 15,545 unedited photos, larger than 600 $\times$ 800 pixels, organized in 5,953 series. To gather the data, the authors of the Automatic Triage for a Photo Series, in a contest like environment, asked the participants to submit personal photo albums. In total, they collected over 350 album submissions from 96 contributors. They then exploit, for time neighbour images, SIFT descriptors [81] and colour similarity to find photo series. Furthermore, for pictures containing human faces, they ensured that each series contain the same group of people. They then split series of photos containing more than 8 shots. In particular, following [125] they adopted a variant of k-means on the 116-dimension global features, where the centre is a representative picture instead of a mean. Finally, all the collected clusters were manually checked to filter clusters with terrible image quality or privacy concerns. After gathering the photo series, they ran a crowd-sourced user study over the Amazon Mechanical Turk (MTurk) [61] to collect human preferences over the best picture of each photos series. Rather than ask participants to rank all the photo series, they instead request to perform pairwise comparisons on images from the same series using a forced-choice methodology in order to better measure small differences. In particular, they asked: "Imagine you take these two photos, and can only keep one. Which one will you choose, and why?". Alongside the choice of the best pictures, participants were asked to describe at least why a particular photo is preferred or why the other photo is not preferred. In the figure 6.3 are reported some of the photos series from the Automatic Triage for a Photo Series dataset. In order to obtain a global ranking for each image series, they used the Bradley-Terry model [11], which describes the probability of choosing the image I_i over I_j as a sigmoid function of the score difference between two photos.

Figure 6.3: Examples of photos series from the Automatic Triage for a Photo Series dataset. In each series of images (*a*, *b*, *c*) is highlighted in green the one preferred by the majority of the people.

Camera Scene Detection Dataset

The Camera Scene Detection Dataset (CamSDD)[105] is a large-scale dataset containing more than 11,000 pictures of sizes 576 × 384 pixels, grouped in 30 of the most important scene categories. Images of the dataset were crawled from Flickr[1] using the same setup as in [49]. After collecting the pictures, images with distorted colours and watermarks, heavily edited pictures or monochrome were manually removed. According to the authors, the dataset was designed to be as much diverse as possible in order to cover all the different environments and shooting conditions. Therefore, each scene category contains images of different places, captured under different viewpoints and rotation angles. The dataset is also balanced with respect to the number of images in each category: on average there are around 350 photos in each group. The 30 categories of the CamSDD dataset, alongside an example of the images representing the classes, are depicted in the figure 6.4.

To cope with the necessity of having a set of images with repeated elements, it has been decided to group the photo series from the Automatic Triage for a Photo Series dataset according to their scene. Since some of the scenes from the CamSDD dataset were not so popular in the images of the considered dataset, they have been removed and take into account only a subset of all the possible scenes. Furthermore, scenes like *cat* and *dog* have been grouped together into the master scene *Animals*. Therefore a total of 19 scenes has been considered, respectively: *Architecture, Backlight, Beach, Blue Sky, Cloudy Sky, Food, Greenery, Group portrait, Indoor, Kids, Mountain, Night shot, Portrait, Snow, Stage concert, Sunset Sunrise, Underwater, Waterfall* and *Animals*.

In order to divide the series into 19 scenes, it has been implemented a majority vote strategy, where the assigned scene of a series is given by the majority of the scene predicted

[1] https://www.flickr.com/

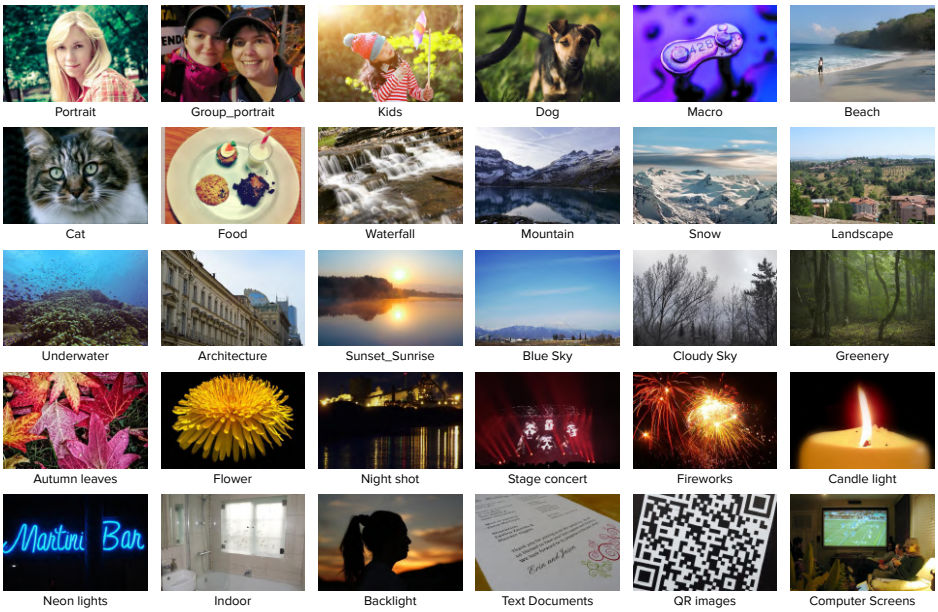


Figure 6.4: The 30 categories of the Camera Scene Detection Dataset.

belonging to that series. To further improve the quality of the scene classification only the labels which had a probability higher than 0.7 have been considered, thus 55% of images have been removed. An overview of the scene distribution is reported in the table 6.1. *Architecture* and *Group_portrait* are the most commons scenes, while *Waterfall* and *Mountain* are the least common ones.

6.4.2 Evaluation metrics

The performances of the proposed method are evaluated over the *Diversity score*, the *Selection precision*, and the *Average probability*.

Due to the nature of the dataset, images belonging to the same series are very similar, therefore the diversity of the methods can be measured by counting the number of different series in the selection, over the number of images to be selected.

The goodness of the selection is evaluated counting the number of *best* images with respect to the number of images to be selected. For each series, the *best* image is the one with the maximum probability of being selected.

Since the Automatic Triage for a Photo Series is human-annotated, and therefore each image is provided with the probability of choosing one particular image rather than the others. The average of the probability values of the selected images (*Average probability*) has been taken into account in support of the *Selection precision* to measure the quality of the results. In fact, it is not entirely taken for granted that there is always one single image better than the others, therefore the probabilities may be similar among different images.

Table 6.1: For each of the 19 scenes, it is reported the number of images belonging to that scene, the number of series, and the average series length.

Scene name	**Number of images**	**Number of series**	**Average series length**
Architecture	1131	443	2.55 ± 0.95
Backlight	129	48	2.69 ± 1.08
Beach	295	110	2.68 ± 0.89
Blue_Sky	209	71	2.94 ± 1.39
Cloudy_Sky	140	57	2.46 ± 0.75
Food	367	164	2.24 ± 0.50
Greenery	164	72	2.28 ± 0.51
Group_portrait	1006	409	2.46 ± 0.82
Indoor	211	90	2.34 ± 0.76
Kids	565	222	2.55 ± 0.89
Mountain	59	23	2.57 ± 1.01
Night_shot	236	87	2.71 ± 1.10
Portrait	71	27	2.63 ± 0.99
Snow	261	93	2.81 ± 1.21
Stage_concert	84	33	2.55 ± 0.70
Sunset_Sunrise	104	36	2.89 ± 0.84
Underwater	88	39	2.26 ± 0.71
Waterfall	57	21	2.71 ± 0.98
Animals	280	100	2.80 ± 1.10
Sum	5457	2145	2.53 ±0.92

Given n series of images $\{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_n\}$, from which we have to select k different images $\{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_k\}$, where $k \leqslant n$.
The *Diversity score* can be defined as:

$$Diversity\ score\ = \frac{\left|\{\mathcal{S}_x | \mathbf{I}_i \in \mathcal{S}_x, 1 \leq i \leq k\}\right|}{k}. \tag{6.1}$$

The *Selection precision* as:

$$Selection\ precision = \frac{\sum_{i=1}^{k} BEST(\mathbf{I}_i)}{k} \tag{6.2}$$

where

$$BEST(\mathbf{I}) = \begin{cases} 1\ if\ p(\mathbf{I})\ = \max_{\mathbf{I}_i \in \mathcal{S}_i\ |\mathbf{I} \in \mathcal{S}_i} p(\mathbf{I}_i) \\ 0\ otherwise \end{cases} \tag{6.3}$$

and $p(\mathbf{I})$ is the probability of the picture $\mathbf{I}$ of being selected.
Finally the *Average probability* can be defined as:

$$Average\ probability\ = \frac{1}{k} \sum_{i=1}^{k} p(\mathbf{I}_i) \tag{6.4}$$

Given that $k \leqslant n$, the *Diversity score* ranges between a value of $1/k$ which reflect a low diversity, and a upper bound of 1, indicating the maximum diversity of the set. *Selection precision* ranges between 0 and 1.

6.4.3 Implementation Details

The proposed framework is tested using Python. The perceived image quality is extracted exploiting the method presented in [70], while the image aesthetics and the *object emphasis* are computed with the system proposed in [69]. K-Means is randomly initialized and executed 10 times in order to select the best results each time and reduce the variability of the results.

The performances of the proposed method are computed over an average of five repetitions. The proposed solution is compared against two different alternative policies, respectively the *Random* and the *Oracle*, and a solution adapted from the state of the arts *High-Contrast Color Sets* [14]. The *Oracle* strategy, given a group of images, selects randomly k images from the best ones. Since for each series of images there is only a single best picture, the *Oracle* will always result to be the best in terms of *Diversity score* and *Selection precision*. The *Random* policy, instead selects k random images from the given group.

The proposed method is also compared with respect to the solution proposed by Campadelli *et al.* . In their work they solve the problem of selecting high-contrast color set as combinatorial optimization problem on graphs. In order to adapt the proposed algorithm to the problem of image summarization, images are considered to be nodes of the graph, like the colours in the reference manuscript, and model the distance between the images with the aforementioned image properties. In particular, the distances between two images is given by the cosine similarity over the ResNeXt-10 features, plus the average of the image quality, the image aesthetics and the *object emphasis*.

Values of k as 5, 10 or 20, have been take into account as a possible number of images that a user would like to extract.

6.5 Results

In this section the proposed method is compared against the aforementioned *Oracle*, *Random* and *High-Contrast Color Sets* policies with respect to the *Diversity score*, the *Selection precision* and the *Average probability*. In the table 6.2 are reported the average performances of the considered policies over five repetitions.

In terms of diversity, as the number of items to be selected increases, the *Diversity score* of the proposed method sightly decreases. From the *Random* perspective, instead, the gap in terms of *Diversity score* start to be noticeable as k increases. This suggests that the proposed method is able to correctly partitioning the images with respect to their content. From this point of view, also *High-Contrast Color Sets* is competitive.

In terms of *Selection precision*, the *Random* strategy fluctuates near the value of 0.3 with a noticeable variance. The proposed method instead reach a *Selection precision* of 0.58 when selecting 5 images, meaning that 3 out of 5 images are the best ones, and decrease to 0.55 when k is equal to 20. The *High-Contrast Color Sets* lies between the *Random* strategy and the proposed method.

The *Average probability* confirms the general behaviour highlighted by the *Selection precision*.

Table 6.2: Average results over 5 repetitions in terms of *Diversity score* and *Selection precision*. Due to the time complexity, the policy *High-Contrast Color Sets* is executed a single time.

Diversity score			
Policy name\k	**5**	**10**	**20**
Random	0.97 ± 0.01	0.95 ± 0.01	0.88 ± 0.01
Oracle	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
High-Contrast Color Sets	1	0.99	0.98
Proposed method	1.0 ± 0.004	0.98 ± 0.01	0.97 ± 0.01
Selection precision			
Policy name\k	5	10	20
Random	0.37 ± 0.05	0.41 ± 0.04	0.40 ± 0.02
Oracle	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
High-Contrast Color Sets	0.45	0.46	0.45
Proposed method	0.58 ± 0.016	0.56 ± 0.01	0.55 ± 0.01
Average probability			
Policy name\k	5	10	20
Random	0.37 ± 0.03	0.41 ± 0.02	0.43 ± 0.01
Oracle	0.69 ± 0.01	0.70 ± 0.01	0.69 ± 0.01
High-Contrast Color Sets	0.45	0.44	0.45
Proposed method	0.50 ± 0.01	0.50 ± 0.01	0.50 ± 0.01

In the figure 6.5, for each of the considered scene category, it is reported the average performances with respect to the *Selection precision* and the *Average probability*. The best performing class result to be *snow*, while the worst one is *cloudy_sky*. Except for the *Mountain* scene, the figure 6.5 also highlight that for scenes with a *Selection precision* less than 0.5 (e.g. *Blue_sky*, *Cloudy_sky*, *Night_shot* and *Animals*) the low values of the *Selection precision* are not reflected by the *Average probability*, suggesting that the selected photos may not be the worst ones. This behaviour is pointed out by the figure 6.6: where the proposed method is wrong it selects the second best picture rather the the best one.

6.5.1 Subjective results

The ground truth of the Automatic Triage for a Photo Series dataset is strictly related to the series of images. Therefore the just presented results only reflect the property of the

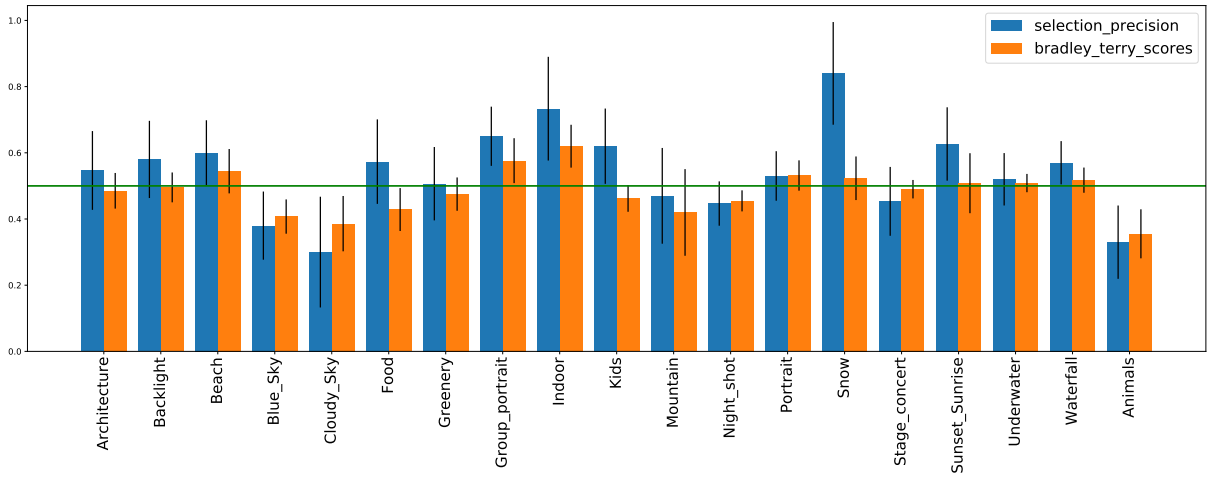


Figure 6.5: Per class average *Selection precision* and *Average probability* of the proposed method. Green line indicate the value of 0.5.

presented method to choose the best images from multiple series and whether that they are different or not. Nonetheless, with the available ground truth, it is not possible to evaluate the goodness of the overall selection. To this end it has been decided to asses the proposed method over the judgment of 7 different human subjects. In that regard, human raters have been asked to select the best image selection with over the three algorithm, respectively the proposed method, the *Oracle* and the *Random* policies. All the 19 scenes and values of k of 5 an 10 have been taken into account. It has been omitted $k = 20$ due to the large number of images selected and the consequently difficulty of compare the three algorithm. Before the experiments, subjects were asked to select the best group of images taking into account the concept of diversity (i.e. privilege group of images without repeated subjects) and the overall aesthetics and quality of the photos providing them some visual examples.

In the figure 6.7 are reported the percentages of votes obtained by each algorithm over the conducted study. The first pie chart refers to the overall scores while the other two plots are specific to the group cardinality 5 and 10 respectively. As pointed out by the first pie chart the proposed method results to be the most selected one followed by the oracle and then by the random policy. This trend can be attributed to the fact that even if the oracle strategy chooses between the best images, it is not true that the selected images are the best ones. The proposed method instead, guided by the considered perceptual properties, aim to select the most beautiful images among all the pictures. In particular, the proposed selection strategy seems to perform better when selecting 5 images rather than 10.

In the figure 6.8 for each of the 19 camera scenes are reported the percentages of preferences given by the human raters. Almost in every situations the proposed method surpasses other methods. Is interesting comparing these results with the one discussed in the figure 6.5. In particular, if we focus on the scenes category that performs awful

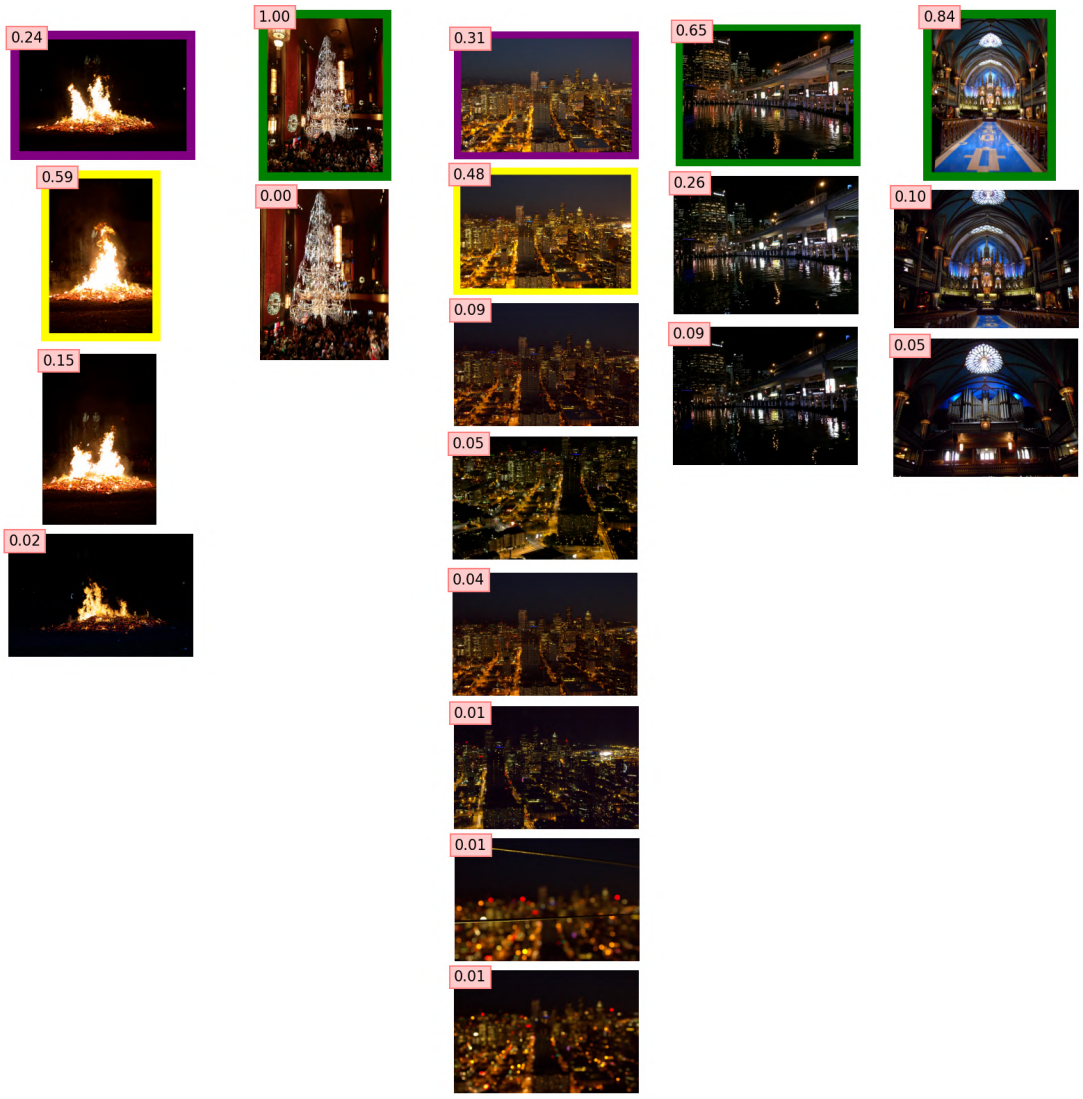


Figure 6.6: Example of the images selected by the proposed method over the scene *Night_shot*. On the first row are reported the images selected by the proposed method while on the others lines are reported the alternatives. The selected images that coincide with the best image from the ground truth are highlighted in green, purple otherwise. Images that are the best one and are different from the chosen one are highlighted in yellow. For each image, in the red box on is reported the ground truth probability of the images.

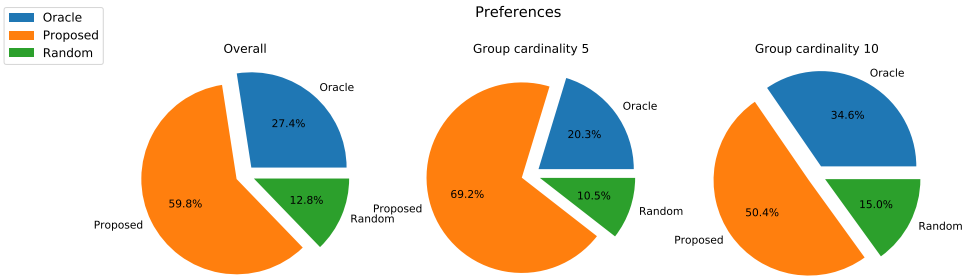


Figure 6.7: Preferences votes distribution over the three considered strategy. The first chart report the overall decision, while the remaining two are with respect to the group cardinality of 5 and 10 respectively,

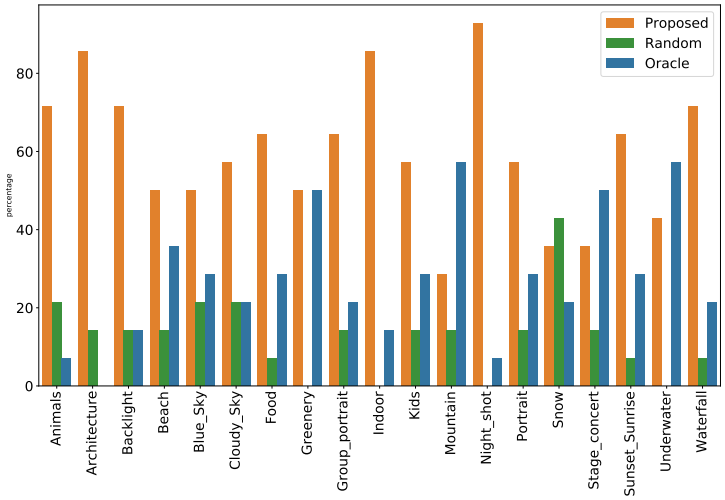


Figure 6.8: For each of the 19 considered scenes, and for each of the considered policies, are reported the percentage of votes given by the human raters.

over the *Selection precision* (less than 0.5), we found that the *Average probability* is more accurate than the *Selection precision*. Therefore only the *Mountain* scene have a lower value of *Selection precision* with respect to the *Average probability*, and in the figure 6.8 the *Mountain* scene is the worst ones. Instead, scenes like *Blue_sky*, *Cloudy_sky*, *Night_shot* and *Animals* where the low values of the *Selection precision* is1 not reflected by the *Average probability*, the proposed algorithm is the most selected one as pointed out by the figure 6.8.

Chapter 7

Conclusions

In this thesis, we have investigated automatic image collection management using convolutional neural networks. To this end, we followed a top-down approach focusing on both the general task and on the single problems that are implied.

The process of summarizing a collection of images has been broken down into single subproblems, namely the *image quality*, *image aesthetics*, *image memorability* and *image diversity*, and the contribution of this thesis are towards each of these problems.

As a first step we addressed image quality assessment by introducing a model that classifies good and bad quality pictures by exploiting the information deep encoded in CNN pretrained on ImageNet, using only a subset of artefact-less images. To this end, we presented a pipeline that relies on the Gram matrix computed over the activation volumes of a CNN to encode the intra-layer correlation. This is further processed used in an anomaly detection fashion to improve the performance obtained by using average intra-class correlation only. The effectiveness of the proposed approach has been demonstrated on three different datasets containing real distorted images. Moreover, cross-dataset experiments conducted highlighted the robustness and generalization skills of the proposed approach in comparison to other algorithms in the literature.

Then we investigated the assessment of image aesthetics and introduced a method based on the prediction of eleven attributes that are closely related to aesthetics judgement. The designed model makes use of MLSP features extracted from an ImageNet-pretrained CNN to predict these eleven attributes with an MLP. Then, an SVR has been trained to infer the aesthetics score of the input images over the prediction of the aforementioned MLP. The experimental results with four different architectures have demonstrated the effectiveness of the proposed approach. In particular, it has been demonstrated that predicting the image aesthetics through related attributes leads to an improvement of 5.5% in terms of SROCC with respect to the state of the art.

The same problem has been tackled with a self-adaptive method that explicitly considers semantic content, style, and composition. In particular, the proposed method exploits the side information related to the aesthetic attributes of an image to build an ad-hoc image aesthetics estimator. The parameters of the aesthetic estimator are adaptively generated from a metamodel consisting of an attribute-conditioned hypernetwork. Given

an image, the resulting model predicts (i) the style and composition of the image, (ii) the distribution of the aesthetic score. Experimental results on three benchmark datasets show that the proposed method achieves comparable performance to previous methods for aesthetic quality classification. Instead, it outperforms state-of-the-art methods for both image aesthetic score regression and aesthetic score distribution prediction.

Subsequently, we explored image memorability and proposed an approach that involves the use of two CNNs respectively trained for image recognition and image memorability. The features extracted from these two CNNs were then used to exploit the knowledge of the context as well as the information about the memorability of the image. Moreover, a soft attention mechanism has been used to focus the model attention on highly informative regions for memorability estimation. The proposed model obtained on the benchmark dataset comparable results with respect to state-of-the-art approaches demonstrating the effectiveness of the proposed method. Moreover, the proposed solution achieved the smallest MSE with the fewest parameters among the methods that outperform the human consistency correlation.

In conclusion, in order to explore the concept of image diversity, a flexible framework that can be used to both explore large scale image datasets and to summarize photo albums has been proposed. The designed pipeline aims at maximizing the diversity and the quality of the resulting set. The proposed method first classifies each image with respect to the camera scene, and then, for each of the scene groups, computes the best k pictures. To this end, the proposed method relies on the features extracted from a ResNeXt-101 32x8d to capture the most diverse pictures, and the concepts of image quality, image aesthetics and whether the image emphasizes foreground objects to select the best pictures. The experimental results over the benchmark dataset demonstrated the effectiveness of the proposed approach. Moreover, experiments conducted over human preferences confirmed the capabilities of the proposed method.

Given the modularity of this thesis to the problem of image collection management, there are several possible ways of extension. Future work will likely focus on other types of perceptual properties. For example, the emotion evoked by an image also plays an important role in the process of selecting images. Therefore being able to automatically estimate such feelings could be beneficial to the overall pipeline. Other peculiar properties of images that can be taken into account to have a better selection of images are the interestingness of the images and the image popularity.

A different direction from exploring other types o perceptual properties extend this thesis is related to the exploitation of these features to the image selection task. To this end, different approaches such as Graph Neural Networks could also be investigated.

Bibliography

[1] Alireza Alaei, Romain Raveaux, and Donatello Conte. Image quality assessment based on regions of interest. *Signal, Image and Video Processing*, 11(4):673–680, 2017.

[2] Seyed Ali Amirshahi, Marius Pedersen, and Stella X Yu. Image quality assessment by comparing cnn features between images. *Electronic Imaging*, 2017(12):42–51, 2017.

[3] Alixandra Barasch, Kristin Diehl, Jackie Silverman, and Gal Zauberman. Photographic memory: The effects of volitional photo taking on memory for visual and auditory aspects of an experience. *Psychological Science*, 28(8):1056–1066, 2017.

[4] Ari Benjamin, David Rolnick, and Konrad Kording. Measuring and regularizing networks in function space. In *International Conference on Learning Representations*, 2018.

[5] Simone Bianco, Remi Cadene, Luigi Celona, and Paolo Napoletano. Benchmark analysis of representative deep neural network architectures. *IEEE Access*, 6:64270–64277, 2018.

[6] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. Predicting image aesthetics with deep learning. In *International Conference on advanced concepts for intelligent vision systems*, pages 117–125. Springer, 2016.

[7] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12(2):355–362, 2018.

[8] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. Disentangling image distortions in deep feature space. *arXiv preprint arXiv:2002.11409*, 2020.

[9] Simone Bianco, Gianluigi Ciocca, Fabrizio Marini, and Raimondo Schettini. Image quality assessment by preprocessing and full reference model combination. In *Image Quality and System Performance VI*, volume 7242, page 72420O. International Society for Optics and Photonics, 2009.

[10] Alan Conrad Bovik. Automatic prediction of perceptual image and video quality. *Proceedings of the IEEE*, 101(9):2008–2024, 2013.

[11] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[12] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, 116:165–178, 2015.

[13] Larry Cahill and James L. McGaugh. A novel demonstration of enhanced memory associated with emotional arousal. *Consciousness and Cognition*, 4(4):410–421, 1995.

[14] Paola Campadelli, Roberto Posenato, and Raimondo Schettini. An algorithm for the selection of high-contrast color sets. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 24(2):132–138, 1999.

[15] Luigi Celona and Raimondo Schettini. Cnn-based image quality assessment of consumer photographs. In *London Imaging Meeting*, volume 2020, pages 129–133. Society for Imaging Science and Technology, 2020.

[16] Luigi Celona and Raimondo Schettini. A genetic algorithm to combine deep features for the aesthetic assessment of images containing faces. *Sensors*, 21(4):1307, 2021.

[17] Andrea Ceroni, Vassilios Solachidis, Claudia Niederée, Olga Papadopoulou, Nattiya Kanhabua, and Vasileios Mezaris. To keep or not to keep: An expectation-oriented photo selection method for personal photo collections. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 187–194, 2015.

[18] Huiwen Chang, Fisher Yu, Jue Wang, Douglas Ashley, and Adam Finkelstein. Automatic triage for a photo series. *ACM Transactions on Graphics (TOG)*, 35(4):1–10, 2016.

[19] Qiuyu Chen, Wei Zhang, Ning Zhou, Peng Lei, Yi Xu, Yu Zheng, and Jianping Fan. Adaptive fractional dilated convolution network for image aesthetics assessment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14114–14123, 2020.

[20] Zhihong Chen. Data covariance learning in aesthetic attributes assessment. *Journal of Applied Mathematics and Physics*, 8(12):2869–2879, 2020.

[21] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995.

[22] Wei-Ta Chu and Yi-Ling Wu. Image style classification based on learnt deep correlation features. *IEEE Transactions on Multimedia*, 20(9):2491–2502, 2018.

[23] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002.

[24] Francis Crick and Christof Koch. Towards a neurobiological theory of consciousness. In *Seminars in the Neurosciences*, volume 2, pages 263–275. Saunders Scientific Publications, 1990.

[25] Gerald C Cupchik. Emotion in aesthetics: Reactive and reflective models. *Poetics*, 23(1-2):177–188, 1995.

[26] Claudio Cusano, Paolo Napoletano, and Raimondo Schettini. Combining multiple features for color texture classification. *Journal of Electronic Imaging*, 25(6):061410, 2016.

[27] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE, 2005.

[28] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision*, pages 288–301. Springer, 2006.

[29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[30] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017.

[31] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR 2011*, pages 1657–1664. IEEE, 2011.

[32] Michael P Eckert and Andrew P Bradley. Perceptual quality metrics applied to still image compression. *Signal processing*, 70(3):177–200, 1998.

[33] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6363–6372. IEEE, 2018.

[34] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020.

[35] Michael Freeman. *The complete guide to light & lighting in digital photography.* Sterling Publishing Company, Inc., 2007.

[36] Michael Freeman. *The photographer's eye: composition and design for better digital photos.* Routledge, 2017.

[37] Fei Gao, Ziyun Li, Jun Yu, Junze Yu, Qingming Huang, and Qi Tian. Style-adaptive photo aesthetic rating via convolutional neural networks and multi-task learning. *Elsevier Neurocomputing*, 395:247–254, 2020.

[38] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in neural information processing systems*, pages 262–270, 2015.

[39] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[40] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015.

[41] Diogo Goncalves, Liweu Liu, and Ana Magalhães. How big can style be? addressing high dimensionality for recommending with style. *arXiv preprint arXiv:1908.10642*, 2019.

[42] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232, 2017.

[43] Jan Hauke and Tomasz Kossowski. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93, 2011.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

[45] Lihuo He, Xinbo Gao, Wen Lu, Xuelong Li, and Dacheng Tao. Image quality assessment based on s-cielab model. *Signal, Image and Video Processing*, 5(3):283–290, 2011.

[46] Yong-Lian Hii, John See, Magzhan Kairanbay, and Lai-Kuan Wong. Multigap: Multi-pooled inception network with text augmentation for aesthetic prediction of photographs. In *International Conference on Image Processing (ICIP)*, pages 1722–1726. IEEE, 2017.

[47] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

[48] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9375–9383, 2019.

[49] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Wespe: weakly supervised photo enhancer for digital cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 691–700, 2018.

[50] Instagram. Our story: A quick walk through our history as a company, 2016.

[51] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*, pages 2429–2437, 2011.

[52] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 145–152. IEEE, 2011.

[53] Johannes Itten. *Design and form: The basic course at the Bauhaus and later*. John Wiley & Sons, 1975.

[54] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[55] Yueying Kao, Ran He, and Kaiqi Huang. Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing*, 26(3):1482–1495, 2017.

[56] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *British Machine Vision Conference (BMVC)*, 2014.

[57] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What makes an image popular? In *International Conference on World Wide Web*, pages 867–876. ACM, 2014.

[58] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *International Conference on Computer Vision (ICCV)*, pages 2390–2398. IEEE, 2015.

[59] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[60] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[61] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456, 2008.

[62] Kodak. *How to take good pictures: a photo guide*. Ballantine Books, 1981.

[63] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision (ECCV)*, 2016.

[64] Talia Konkle, Timothy F Brady, George A Alvarez, and Aude Oliva. Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological Science*, 21(11):1551–1556, 2010.

[65] Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine learning research*, 5(Jan):27–72, 2004.

[66] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178. IEEE, 2006.

[67] Jun-Tae Lee, Han-Ul Kim, Chul Lee, and Chang-Su Kim. Photographic composition classification and dominant geometric element detection for outdoor scenes. *Elsevier Journal of Visual Communication and Image Representation*, 55:91–105, 2018.

[68] Marco Leonardi, Luigi Celona, Paolo Napoletano, Simone Bianco, Raimondo Schettini, Franco Manessi, and Alessandro Rozza. Image memorability using diverse visual features and soft attention. In *International Conference on Image Analysis and Processing*, pages 171–180. Springer, 2019.

[69] Marco Leonardi, Paolo Napoletano, Alessandro Rozza, and Raimondo Schettini. Modeling image aesthetics through aesthetic-related attributes. In *London Imaging Meeting*, volume 2021, pages –. Society for Imaging Science and Technology, 2021.

[70] Marco Leonardi, Paolo Napoletano, Raimondo Schettini, and Alessandro Rozza. No reference, opinion unaware image quality assessment by anomaly detection. *Sensors*, 21(3):994, 2021.

[71] Cheng-Hung Li, Chih-Yi Chiu, Chun-Rong Huang, Chu-Song Chen, and Lee-Feng Chien. Image content clustering and summarization for photo collections. In *2006 IEEE International Conference on Multimedia and Expo*, pages 1033–1036. IEEE, 2006.

[72] Congcong Li, Alexander C Loui, and Tsuhan Chen. Towards aesthetics: A photo quality assessment and photo selection system. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 827–830, 2010.

[73] Qiang Li and Zhou Wang. Reduced-reference image quality assessment using divisive normalization-based image representation. *IEEE journal of selected topics in signal processing*, 3(2):202–211, 2009.

[74] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019.

[75] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Deepfl-iqa: Weak supervision for deep iqa feature learning. *arXiv preprint arXiv:2001.08113*, 2020.

[76] Kwan-Yee Lin and Guanxiang Wang. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 732–741, 2018.

[77] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[78] Dong Liu, Rohit Puri, Nagendra Kamath, and Subhabrata Bhattacharya. Composition-aware image aesthetics assessment. In *Winter Conference on Applications of Computer Vision*, pages 3569–3578, 2020.

[79] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1040–1049, 2017.

[80] Yutao Liu, Guangtao Zhai, Ke Gu, Xianming Liu, Debin Zhao, and Wen Gao. Reduced-reference image quality assessment in free-energy principle and sparse representation. *IEEE Transactions on Multimedia*, 20(2):379–391, 2017.

[81] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[82] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. Rapid: Rating pictorial aesthetics using deep learning. In *International Conference on Multimedia*, pages 457–466. ACM, 2014.

[83] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *International Conference on Computer Vision (ICCV)*, pages 990–998. IEEE, 2015.

[84] Wei Luo, Xiaogang Wang, and Xiaoou Tang. Content-based photo quality assessment. In *International Conference on Computer Vision (ICCV)*, pages 2206–2213. IEEE, 2011.

[85] Mathias Lux, Marian Kogler, and Manfred Del Fabro. Why did you take this photo: a study on user intentions in digital photo productions. In *Proceedings of the 2010 ACM workshop on Social, adaptive and personalized multimedia interaction and access*, pages 41–44, 2010.

[86] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2017.

[87] Shuang Ma, Jing Liu, and Chang Wen Chen. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4535–4544. IEEE, 2017.

[88] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[89] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[90] Long Mai, Hailin Jin, and Feng Liu. Composition-preserving deep photo aesthetics assessment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 497–506. IEEE, 2016.

[91] Gautam Malu, Raju S Bapi, and Bipin Indurkhya. Learning photography aesthetics with deep cnns. *arXiv preprint arXiv:1707.03981*, 2017.

[92] Matei Mancas and Olivier Le Meur. Memorability of natural scenes: The role of attention. In *International Conference on Image Processing (ICIP)*, pages 196–200. IEEE, 2013.

[93] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *International Conference on Computer Vision (ICCV)*, pages 1784–1791. IEEE, 2011.

[94] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364, 2011.

[95] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2408–2415. IEEE, 2012.

[96] Paolo Napoletano, Flavio Piccoli, and Raimondo Schettini. Anomaly detection in nanofibrous materials by cnn-based self-similarity. *Sensors*, 18(1):209, 2018.

[97] Pere Obrador, Ludwig Schmidt-Hackenberg, and Nuria Oliver. The role of image composition in image aesthetics. In *International Conference on Image Processing*, pages 3185–3188. IEEE, 2010.

[98] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Springer International Journal of Computer Vision*, 42(3):145–175, 2001.

[99] Stephen E Palmer, Karen B Schloss, and Jonathan Sammartino. Visual aesthetics and human preference. *Annual review of psychology*, 64:77–107, 2013.

[100] Bowen Pan, Shangfei Wang, and Qisheng Jiang. Image aesthetic assessment assisted by attributes through adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 679–686, 2019.

[101] Thrasyvoulos N Pappas, Robert J Safranek, and Junqing Chen. Perceptual criteria for image quality evaluation. *Handbook of image and video processing*, 110, 2000.

[102] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[103] W Pirie. *Spearman Rank Correlation Coefficient*, volume 8. 08 2006.

[104] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[105] Angeline Pouget, Sidharth Ramesh, Maximilian Giang, Ramithan Chandrapalan, Toni Tanner, Moritz Prussing, Radu Timofte, and Andrey Ignatov. Fast and accurate camera scene detection on smartphones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2569–2580, 2021.

[106] Gajjala Viswanatha Reddy, Snehasis Mukherjee, and Mainak Thakur. Measuring photography aesthetics with deep cnns. *IET Image Processing*, 14(8):1561–1570, 2020.

[107] Abdul Rehman and Zhou Wang. Reduced-reference image quality assessment by structural similarity estimation. *IEEE transactions on image processing*, 21(8):3378–3389, 2012.

[108] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Springer International Journal of Computer Vision*, 115(3):211–252, 2015.

[109] Pakizar Shamoi, Atsushi Inoue, and Hiroharu Kawanaka. Modeling aesthetic preferences: Color coordination and fuzzy sets. *Elsevier Fuzzy Sets and Systems*, 395:217–234, 2020.

[110] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 3. Minneapolis, MN, 2007.

[111] Yangyang Shu, Qian Li, Shaowu Liu, and Guandong Xu. Learning with privileged information for photo aesthetic assessment. *Neurocomputing*, 404:304–316, 2020.

[112] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. How to make an image more memorable?: A deep style transfer approach. In *International Conference on Multimedia Retrieval (ICMR)*, pages 322–329. ACM, 2017.

[113] Gunnar A. Sigurdsson, Xinlei Chen, and Abhinav Gupta. Learning visual storylines with skipping recurrent neural networks. *European Conference on Computer Vision*, 2016.

[114] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[115] Pinaki Sinha, Sharad Mehrotra, and Ramesh Jain. Summarization of personal photologs using multidimensional content and context. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, pages 1–8, 2011.

[116] Hammad Squalli-Houssaini, Ngoc QK Duong, Marquant Gwenaëlle, and Claire-Hélène Demarty. Deep learning for predicting image memorability. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2371–2375. IEEE, 2018.

[117] Lionel Standing. Learning 10000 pictures. *The Quarterly journal of experimental psychology*, 25(2):207–222, 1973.

[118] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

[119] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[120] Xiaoou Tang, Wei Luo, and Xiaogang Wang. Content-based photo quality assessment. *IEEE Transactions on Multimedia*, 15(8):1930–1943, 2013.

[121] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[122] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[123] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Elsevier Neural networks*, 22(5-6):544–557, 2009.

[124] Domonkos Varga, Dietmar Saupe, and Tamás Szirányi. Deeprn: A content preserving deep architecture for blind image quality assessment. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018.

[125] Xin-Jing Wang, Lei Zhang, and Ce Liu. Duplicate discovery on 2 billion internet images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 429–436, 2013.

[126] Yufei Wang, Zhe Lin, Xiaohui Shen, Radomir Mech, Gavin Miller, and Garrison W Cottrell. Event-specific image importance. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4810–4819, 2016.

[127] Yufei Wang, Zhe Lin, Xiaohui Shen, Radomír Mech, Gavin Miller, and Garrison W Cottrell. Recognizing and curating photo albums via event-specific image importance. *arXiv preprint arXiv:1707.05911*, 2017.

[128] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[129] Zhou Wang and Eero P Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *Human Vision and Electronic Imaging X*, volume 5666, pages 149–159. International Society for Optics and Photonics, 2005.

[130] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[131] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[132] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, pages 2048–2057, 2015.

[133] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1098–1105. IEEE, 2012.

[134] Che-Hua Yeh, Yuan-Chen Ho, Brian A Barsky, and Ming Ouhyoung. Personalized photograph ranking and selection system. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 211–220, 2010.

[135] Zephoria. The top 20 valuable facebook statistics, 2015.

[136] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.

[137] Luming Zhang, Yue Gao, Roger Zimmermann, Qi Tian, and Xuelong Li. Fusion of multichannel local and global structural cues for photo aesthetics evaluation. *IEEE Transactions on Image Processing*, 23(3):1419–1429, 2014.

[138] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[139] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[140] Xiaodan Zhang, Xinbo Gao, Lihuo He, and Wen Lu. Mscan: Multimodal self-and-collaborative attention network for image aesthetic prediction tasks. *Elsevier Neurocomputing*, 430:14–23, 2021.

[141] Xiaodan Zhang, Xinbo Gao, Wen Lu, and Lihuo He. A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction. *IEEE Transactions on Multimedia*, 21(11):2815–2826, 2019.

[142] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.

[143] Ye Zhou, Xin Lu, Junping Zhang, and James Z Wang. Joint image and text representation for aesthetics analysis. In *International Conference on Multimedia*, pages 262–266. ACM, 2016.

Appendices

Appendix A

Human subjectivity

Classical features, intended as measurable attributes of an object, such as its colour, form or orientation have usually been defined by clear-cut criteria: i.e. for the image composition rule of thirds, the subject of the photo must be placed along with the imaginary guidelines that divide the image into nine equal parts. This is not the case for perceptual features: i.e. the judgment of aesthetics quality of picture can be affected by commonly established photographic rules, such as image composition, lighting [35] and contrast [53]. But there are no strict rules that define if a photo is aesthetically pleasing or not with respect to these attributes. Not all images with the rule of thirds can be considered good photos. It should be evident that people differ in their preferences, which depends on highly subjective factors not easily describable. Nonetheless, it is possible to capture the subjectivity over perceptual properties by collecting a large amount of human perceived scores. One of the most reliable and definitive ways of gathering subjective human evaluations is adopting an online crowdsourcing platform where users can assess the evaluation virtually. One of the most popular systems, employed to gather a large number of opinions from a diverse distributed populace, is Amazon Mechanical Turk (AMT)[1]. Given a set of images that have to be labelled, participants to the task are asked to provide an opinion on the perceived visual attribute (i.e. quality, aesthetics) of the presented images. To ensure adequate quality of the answers experiments are designed with an entry barrier (e.g. pass a qualification test) and incentive for participation using a micropayment scheme.

Once the data have been collected, in order to define the ground-truth for designing and evaluating reliable models, a "global opinion" is usually computed for each image in the set. According to the number of both participants and labels, there are possible ways to obtain a "global opinion", such as unanimous agreement and Mean Opinion Score (MOS). Typically when the number of opinions assigned to each image is limited, and the label space is narrow (e.g. in the CUHK-PQ [120] dataset for the image quality each image is labelled by ten participants and the possible classes are two) is used the unanimous agreement. In particular, each image is assigned to the label that all the

[1] https://www.mturk.com/

participants agreed (e.g. a photo is classified as high or low quality only if eight out of the ten reviewers agree on its assessment). When hundred of opinions per image are available, and the associated rating varies on a fixed integer scale (i.e. in the AVA dataset each image has an average of 210 votes and a score ranging from 1 to 10 [95]) the average score for the image I, also known as MOS, is commonly used as a proxy for the perceived image quality or the image aesthetics. Given the image I, the MOS can be defined as:

$$MOS(I) = \frac{1}{N}\sum_{i=1}^{N} r_i(I), \tag{A.1}$$

where $r_i(I)$ is the $i-th$ individual score of the image I.

A.1 Goodness of the fit

The ambition of subjective property assessment models (i.e. image quality, aesthetic and memorability assessment) is to emulate the human perception, consequently to obtain a high correlation with the MOS. The most commonly used metrics to evaluate the performance of subjective property assessment methods are respectively the Pearson's Linear Correlation Coefficient (PLCC) and the Spearman's Rank-order Correlation Coefficient (SROCC). Both are used to compare the scores predicted by the models and the subjective opinion scores provided by the dataset. The PLCC correlation evaluates the linear relationship between two continuous variables while the SROCC evaluates the monotonic relationship between two continuous or ordinal variables. Both indexes are used to evaluate the correlation between variables. However, according to [43], Spearman's index is more suitable to evaluate relationships involving ordinal variables than Pearson's index. The automatic image quality assessment methods aim to predict a quality index of the image which simulates the Mean Opinion Score achieved thanks to subjective evaluation of users. Since both the quality index and the MOS are continuous and ordinal variables, the SROCC is the most reliable evaluation metric for IQA methods.

The PLCC is a statistic that captures the linear correlation between the predicted scores and MOS; it ranges between $+1$ and -1, where a value of $+1$ or -1 reflects a totally positive, or negative respectively, linear correlation, and a value of 0 is no linear correlation. Given n as the number of the considered samples, x_i and y_i the sample points indexed with i, $\bar{x}$ and $\bar{y}$ the means of each sample distribution; we can define the PLCC as follows:

$$PLCC = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}. \tag{A.2}$$

Instead, the SROCC operates on the rank of the data points ignoring the relative distances between them, hence assesses the monotonic relationships between the actual and predicted scores. As the PLCC, it varies in the interval $[+1,-1]$ and for n considered

samples it is defined as follow:

$$SROCC = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n\,(n^2 - 1)}, \tag{A.3}$$

where $d_i = (rank(x_i) - rank(y_i)$ is the difference between the two ranks of each sample.

Alongside the PLCC and SROCC, the quality of a regression model is commonly evaluated comparing the estimated outputs against the actual values, in particular measuring the difference, or error, between them. Some of the most popular error metrics are the Mean Absolute Error (MAE), the Mean Squared Error (MSE), and the Root Mean Squared Error (RMSE).

The MAE describe how close the estimated values are to the real ones. It is defined as follow:

$$MAE(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|, \tag{A.4}$$

where y_i are the output values, $\hat{y}_i$ are the ground-truth values, and N is the number of samples. It describes the magnitude of the residuals, and because of the absolute value, it does not indicate underperformance or overperformance of the model. MAE range from 0 to $+\infty$ and smaller values indicate better results.

The MSE is similar to the MAE except that the differences are squared. The equation can be formalized as:

$$MSE(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2, \tag{A.5}$$

where again $\hat{y}_i$ are the actual scores, y_i the predicted scores and N is the number of samples. As the MAE it range between 0 and $+\infty$, where 0 indicate that the model is a perfect predictor. The main difference between MSE and MAE is that MSE penalized more the outliers than the MAE because of the square.

Finally RMSE is basically the MSE scaled to the original error:

$$RMSE(\hat{y}, y) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2}. \tag{A.6}$$