Automatic analysis of image quality criteria in natural scenes using deep neural networks

Marcelin Tworski

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAT031

Thèse de doctorat

Automatic analysis of image quality criteria in natural scenes using deep neural networks

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 Ecole Doctorale de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Signal, image, automatique et robotique

Thèse présentée et soutenue à Palaiseau, le 14 Décembre 2023, par

MARCELIN TWORSKI

Composition du Jury :

Anissa Mokraoui Professeure des universités, Université Sorbonne Paris Nord (Laboratoire de Traitement et Transport de l'Information)	Présidente du jury Examinatrice
Frédéric Dufaux Directeur de recherche, CentraleSupélec (Laboratoire des Signaux et Systèmes)	Rapporteur
Pietro Zanuttigh Associate Professor, University of Padova (Laboratorio di Tecnologia e Telecomunicazioni Multimediali)	Rapporteur
Yann Gousseau Professeur, Télécom Paris (IMAGES)	Examinateur
Marco Cagnazzo Professeur, Télécom Paris (Image, Données, Signal)	Directeur de thèse
Stéphane Lathuilière Maitre de conférences, Télécom Paris (Image, Données, Signal)	Co-directeur de thèse

Abstract

As smartphone cameras became more prevalent than traditional camera systems, the demand for precise measurements increased. This Ph.D. dissertation proposes using deep learning systems to evaluate image-quality criteria specific to smartphone camera evaluation, and more specifically texture evaluation. This dissertation addresses several limitations in current image-quality assessment methods for smartphone cameras. Deep learning systems struggle with computational complexity due to high-resolution smartphone images, and downsizing would lead to information loss for evaluating noise and details preservation. Consequently, it is essential to find the relevant image regions to assess a camera attribute to alleviate these problems. Additionally, the lack of suitable datasets hinders the development of learning-based methods aimed at benchmarking smartphone cameras. Furthermore, when comparing cameras, it is essential to capture the same content to facilitate direct comparison. In standard camera benchmarking protocols, multiple shots are collected from the same content. This setting deviates from traditional machine learning approaches, where training and test data are assumed to be independent and identically distributed (*iid*). However, the non-independent nature of our data is frequently overlooked in the image-quality assessment literature.

To overcome these challenges, this research introduces several contributions: **(i)** A region selection method is introduced to automatically detect relevant regions for evaluating specific quality attributes. Adapting the class activation map method for a regression problem, we outperform traditional chart-based approaches in evaluating texture quality and permitting the usage of deep learning methods on charts shot in laboratory conditions. In this work, we use texture quality as an illustrative example of camera quality attributes. However, our methodology is designed to be applicable to other attributes, such as noise, as well. **(ii)** A new in-the-wild dataset is created to accurately reflect the complex mixture of defects commonly found in smartphone camera images and reflect the scenario of camera benchmarking, where several different scenes are shot by multiple camera devices. This dataset, annotated through pairwise comparisons, allows us to perform a large evaluation of different methods in different practical scenarios, setting guidelines for the usage of deep learning systems for camera quality evaluation. **(iii)** We introduce a new image quality assessment setup and method where we go beyond the traditional *iid* assumption. We consider multiple images with varying quality of the same content available at test time. We use the specificity of this camera quality estimation setting to enhance the quality prediction accuracy by introducing a batch-based pseudo-reference which allows us to use full-reference methods in the no-reference setting.

Résumé

Alors que les caméras de smartphones sont devenues plus répandues que les systèmes de caméras traditionnels, la demande de mesures précises de qualité a augmenté. Cette thèse de doctorat propose d'utiliser des systèmes d'apprentissage profond pour évaluer de manière plus approfondie les critères de qualité d'image spécifiques à l'évaluation des caméras de smartphones, et plus particulièrement l'évaluation de la texture, là où les approches précédentes reposaient sur la photographie de mires en laboratoire ou bien n'évaluaient qu'une qualité globale de l'image. Cette thèse aborde de manière exhaustive plusieurs limites des méthodes actuelles d'évaluation de la qualité d'image pour les caméras de smartphones. Les systèmes d'apprentissage profond se heurtent à une complexité de calcul accrue due à la haute résolution des images de smartphones, et une réduction de la taille entraînerait inévitablement une perte d'informations pour apprécier pleinement la préservation du bruit et des détails. Par conséquent, il est essentiel de développer des stratégies pour identifier les régions d'image pertinentes pour évaluer les attributs d'un appareil photo afin d'atténuer ces problèmes. En outre, le manque d'ensembles de données appropriés constitue un obstacle majeur qui entrave le développement de méthodes basées sur l'apprentissage visant à évaluer de manière robuste les appareils photo de smartphones. Par ailleurs, lors de la comparaison des appareils photo, il est essentiel de photographier le même contenu pour faciliter la comparaison directe et garantir une comparaison juste et équitable. Dans les protocoles standard d'évaluation comparative des appareils photo, plusieurs photos sont prises à partir du même contenu. Ce cadre méthodologique s'écarte des approches traditionnelles d'apprentissage automatique, dans lesquelles les données d'apprentissage et de test sont supposées être indépendantes et identiquement distribuées (*iid*). Cependant, la nature non indépendante de nos données est souvent négligée dans la littérature sur l'évaluation de la qualité des images.

Pour relever ces défis, cette recherche apporte plusieurs contributions : **(i)**. Une méthode de sélection des régions est introduite pour détecter automatiquement les régions pertinentes pour l'évaluation d'attributs de qualité spécifiques. En adaptant la méthode de la carte d'activation de classe à un problème de régression, nous sommes capable de calculer en utilisant plusieurs images d'entrainement une index de pertinence pour chaque zone de l'image. Avec cette nouvel information, nous sommes capable de surpasser les approches traditionnelles basées sur des mires spécialisées pour évaluer la qualité de la texture et permettre l'utilisation de méthodes d'apprentissage profond sur des mires photographiées en laboratoire. Dans ce travail, nous utilisons la qualité de la texture comme un exemple des attributs de qualité de la caméra. Cependant, notre méthodologie est conçue pour

s'appliquer également à d'autres attributs, tels que le bruit, ce que nous testons expérimentalement. **(ii)** Un nouvel ensemble de données in-the-wild est créé pour refléter avec précision le mélange complexe de défauts que l'on trouve couramment dans les images d'appareils photo de smartphones et pour refléter le scénario de l'évaluation comparative des appareils photo, dans lequel plusieurs scènes différentes sont photographiées par plusieurs appareils photo. Cet ensemble de données est annoté par des comparaisons deux à deux en maximisant l'information apporté par une nouvelle comparaison considérant les comparaisons précédemment effectués. Il nous permet d'effectuer une large évaluation de différentes méthodes dans différents scénarios pratiques, établissant des lignes directrices pour l'utilisation de systèmes d'apprentissage profond pour l'évaluation de la qualité des appareils photo. **(iii)** Nous introduisons une nouvelle configuration et une nouvelle méthode d'évaluation de la qualité des images qui vont au-delà de l'hypothèse *iid* traditionnelle. Nous considérons plusieurs images de qualité variable du même contenu disponibles au moment du test. Nous utilisons la spécificité de ce cadre d'estimation de la qualité de la caméra pour améliorer la précision de la prédiction de la qualité en introduisant une pseudo-référence calculée à partir des images présentes dans le batch. Cette méthode nous permet d'utiliser des méthodes full-reference dans un cadre sans référence. Cette approche novatrice nous permets d'effectuer une adaptation du système d'évaluations à un nouveau domaine, ici une nouvelle scène photographiée, entièrement au moment du test. Ainsi, cette thèse de doctorat suit une progression vers l'évaluation automatique sur scènes naturelles. Nous commençons sur des contenus naturels en laboratoire, et utilisons les méthodes d'apprentissage profond. Ensuite, sur des scènes naturelles, nous constituons une base de données adaptée à notre problématique et montrons l'utilité de l'emploi d'une référence de haute qualité. Enfin nous proposons une nouvelle approche pour pouvoir nous passer de cette référence de haute qualité tout en maintenant de hautes performances.

Acknoledgements/Remerciements

Avant toute chose, j'aimerais remercier très chaleureusement Stéphane, qui a permis que cette thèse soit possible. Je n'aurais pas pu demander un meilleur encadrement. Il a su toujours proposer de nouvelles pistes à explorer lorsque j'étais démotivé. Ensuite, j'aimerais remercier Marco, qui malgré la distance, a continué de s'intéresser à l'évolution de ma thèse. J'aimerais aussi remercier Attilio et Claudio, qui avec Marco ont accepté et permis que je fasse cette thèse. Ensuite, j'aimerais remercier mes collègues à DXOMARK, dont Benoit qui m'a beaucoup soutenu, ainsi que mes collègues de tous les jours avec qui discuter a pu m'apporter, Salim, Nicolas, Théo et Ana. Egalement, je remercie mes collègues de laboratoire.

Plus personnellement, j'aimerais remercier ma mère, qui a toujours tout fait pour moi et mes frères, Edouard et Jacques que je remercie aussi. Je remercie mon père qui a participé à me donner le goût des sciences. J'aimerais remercier mes amis, Aurélien, Théo Marie, dont leur amitié a toujours été d'une grande aide psychologique. J'aimerais remercier Laure qui m'a toujours remotivé et toujours dit que le plus important était ma thèse. Également, je remercie mes colocataires et amis pour la plus grande durée de cette thèse, Benoit, Matthieu, Quentin et Ugo.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Scientific and technological context

1.1.1 Rise of Smartphone Photography

Fifteen years ago, most photographers were using either compact or DSLR cameras. Today, the vast majority of pictures are taken with smartphones. The transition began around 2011, with over 25% of photographs [4] taken using smartphones. By 2015, over a trillion photos were being captured annually, mostly on smartphones. The rise of smartphone photography is evident in their widespread use and increasing market share, outpacing sales of traditional digital cameras tenfold by 2013. Despite catching many camera manufacturers off guard, the ease and convenience of using a smartphone for photography quickly became apparent. Handling photos taken on a traditional digital camera is often a complicated and time-consuming process that involves several steps, including setup, capture, transfer, processing, and publishing. The integration of smartphones into daily life made them available for users, and the streamlined photo sharing process through cloud connectivity made it the preferred choice for casual photography. Photographs taken with a smartphone are usually spontaneous and taken with little or no prior setup, using the camera app's default settings. Minimal post-processing is typically required, enabling fast sharing with friends and family.

The quick gain in popularity of smartphone photography sparked greater interest and higher demands for high-quality photos, leading smartphone manufacturers to prioritize improving their cameras and image processing capabilities. In many instances, it is now difficult to distinguish between a photo taken with a smartphone and one captured with a professional full-frame camera. The traditional tell-tale signs of smartphone photography are no longer reliable indicators. Smartphone camera performance can be evaluated along several dimensions and hundreds of attributes, and many image quality problems can be easily corrected through automated processing. This includes fixing lens shading, optical distortion, poor tonal range, and chromatic aberration. Such correction requires precise measurement of the camera system's optic and sensor combination, which smartphone makers can accomplish as they provide the complete system including the sensor, optics, and image processing. However traditional cameras still have an edge due to their sensor size compared to smartphones which suffer both in detail preservation and noise, since the level of noise in an image is proportional to the amount of light captured.

The camera and its image quality became a critical selling point for a smartphone. As a result, manufacturers began to invest significantly in reducing this gap. They started by using larger image sensors with higher resolutions and improving image capture and processing. From 2000 to 2008, smartphone sensor resolution increased by more than ten times. Surprisingly, the sensors increased resolution and sensitivity alone was only a small part of the improvement in terms of image quality; the increased processing power of mobile devices, of around 100 times, along with new algorithms, played a much larger role in this. Indeed, new denoising algorithms were proposed, such as BM3D [30], WNNM [48] or TWSC [138] which is tailored for real-world images. Furthermore, to reduce noise in images, smartphone makers began to stack multiple captures utilizing computational imaging. These improvements led to smartphones overtaking compact cameras for many uses, and even achieving better performances than older DSLR [51]. Even more surprisingly, smartphones achieve a wider dynamic range than DSLR, even though the sensor is way smaller. Indeed full-frame cameras capture scenes in a single frame, which is perfect for well-lit scenes but falls short in HDR, while smartphones stack multiple images with techniques such as HDR+ [50]. A

DSLR user would need to shoot multiple images from the HDR scenes in RAW and utilize post-processing software. Furthermore, manufacturers began to incorporate multiple camera modules on their phones, as each module is fairly small. These multiple cameras modules have been used by smartphone makers to provide optical zoom, specialized shooting modes such as black and white photography, and bokeh effects [105, 51]. Nowadays, flagship phones have up to five camera modules, a far cry from the single main camera module they used to have.

1.1.2 Development of Deep-Learning based Algorithms

Originally proposed by Le Cun in 1998 [75], the now widely used concept of convolutional neural networks, was first useful for only a few tasks, such as handwritten character recognition. However, due to the democratization of the internet, larger amounts of data became available, and development in GPU technologies greatly improved processing power. These evolutions led to more training possibilities and more complex convolutional architecture. These evolutions allowed for a wider range of applications of deep learning. The ImageNet image classification challenge [31], has been the first breakthrough of these technologies. In 2011, the contest was dominated by hand-engineered computer vision features and had a top-5 error rate of over 25%. However, in 2012, the convolutional neural network AlexNet [71] won the contest with a significant reduction in the error rate to 16%. In 2023, the top-5 error-rate hovers around only 1% on this benchmark. Classical vision tasks such as semantic segmentation, object detection [43, 42, 61, 52] and classification, upscaling [76, 133] are all dominated by deep learning techniques. Deep learning algorithms made possible various image treatment tasks, such as compression [8], inpainting, style transfer [83], automatic captioning [103], or text-to-image generation [106]. Furthermore, these advances have implications for various fields, such as medical imaging diagnostic [5], protein folding [64], financial fraud detection [21] or language translation [100].

1.2 Context of the thesis

This thesis had the particularity of being an industrial agreement for formation through research (CIFRE). Concretely, I was working with DXOMARK, a medium-sized technological company, that specialized in benchmarking smartphones, cameras, or audio devices. For my academic affiliation, I was integrated in the Multimedia (MM) team of the Laboratoire de Traitement et Communication de l'Information (LTCI), a laboratory of Télécom Paris, which is part of the Institut Polytechnique de Paris (IPP).

1.2.1 DXOMARK Activity and Protocol

The area of expertise for my unit was image quality evaluation. DXOMARK aims to provide user-centered evaluation benchmarks of the real-world performance of cameras, usually cameras integrated in smartphones. They also offer consulting services to help manufacturers tune and optimize the device throughout the camera development cycle. Indeed, the configuration between different optical systems with various image enhancement software is not trivial. In order to find the best trade-off in terms of image quality, it requires a thorough evaluation of the device according to various camera and image quality attributes as well as a performance breakdown in various situations. Both photo and video modes are tested. However, in the rest of this protocol presentation, as well as in the rest of the manuscript, we will focus on the evaluation of still images. In the DXOMARK protocol, the evaluation of the cameras is broken down into 6 different attributes, which are further broken down into sub-attributes in various conditions.

- **Exposure** Evaluation of the camera's ability to accurately adjust and capture brightness in both the subject and background. This is testing if the exposure value is appropriate. For portraits, faces exposure is emphasized. Contrast evaluation is also a part of this attribute, as well as dynamic range.
- **Color** Evaluation of the camera's ability to reproduce color accurately under various lighting and scenarios, with a focus on pleasing color rendering for viewers, and in

consequence, there is some tolerance with saturation. For portrait, an emphasis is added on skin tone rendering (ranging from deep to fair to light). White balance, which refers to the camera's ability to adjust its color temperature to match the lighting conditions and produce accurate colors in the image, is also evaluated.

- **Autofocus** Evaluation of the camera's speed and accuracy of focusing on subjects in different lighting conditions.
- **Texture** Evaluation of the camera's ability to capture and preserve small, intricate details, such as those found on object surfaces. As explained in 1.1, this ability is far from depending only on the sensor's resolution.
- **Noise** Evaluation of noise levels in an image, particularly in low light conditions. Different aspects of the noise in an image are evaluated, such as its spatial correlation, its chromaticity as well as its frequency (coarse vs fine-grained). Usually, a fine-grained white noise in luminance is not a drawback of an image if its intensity is not too large.
- **Artifacts** Evaluation of unintended visual anomalies, such as distortion, halo effects, color fringing, or loss of sharpness on the image edges, which appear in an image and are not present in the original object being imaged.

To evaluate the various attributes of a single device, it is necessary to assess it under a wide range of conditions. This requires a significant amount of data, with over 1500 images needed to accurately gauge its performance. To ensure consistent and precise results, a large portion of these tests are conducted in a laboratory setting. This provides a controlled environment where conditions can be easily repeated, and specialized lighting equipment can be used to provide optimal illumination. The subjects of these photos are typically standardized charts. These charts have been specifically designed to provide a consistent and repeatable way of measuring the performance of the device under test. While these laboratory tests can provide evaluations of all attributes necessary to assess the camera quality, it is preferable to conduct a large portion of the tests in a non-laboratory environment since there are several limitations to using only lab evaluation. First, the chart used is often industry-wide standards,

and the manufacturer probably already used them to fine-tune the device on them. It may in this case not reflect so well the actual capabilities of the camera. Second, as discussed in 1.1, the final rendering is the result of an imaging pipeline that is heavily reliant on software, and often the behavior of these algorithms is non-linear and thus content dependent. (For traditional cameras, the results of laboratory tests are closer to the actual capabilities of the device). In this context, it is critical to perform some of these tests on content that are likely to correspond to content photographed by end-users. However, for these tests, we do not have the same repeatability as in the laboratory. This problem is usually alleviated by shooting the device under test along with a previously evaluated camera to capture images in the same conditions. Furthermore, since the photographed content is not a standardized chart, it usually requires human evaluation of the different quality attributes.

1.2.2 Limitations of the literature regarding the DXOMARK protocol

Evaluating numerous images based on multiple attributes, which includes multiple camera quality criteria, is a time-consuming task for qualified image quality analysts. With this Ph.D. project, DXOMARK aimed to reduce the evaluation time and automate certain criteria evaluation. However, traditional methods are not suitable for this evaluation as the image content can vary and is not designed for automatic quality assessment. Therefore, considering the increasing applications of learning-based algorithms, we decided to resort to using machine learning techniques to evaluate the image quality criteria on these images.

There is a vast amount of literature available on image quality assessment (see Sections 2.3 and 2.4), but it has several limitations when it comes to addressing our specific problem:

- Smartphone image resolutions are generally high, often larger than ten megapixels. The large size of images presents a challenge for current IQA models, which tend to have high computational complexities. For example, the SPAQ dataset downsampled their smartphone photography dataset to images with a shorter dimension of 512 pixels. This is adequate for the evaluation of many camera quality attributes, such as exposure or color, but deletes information regarding noise or details preservation. On the

other hand, not all regions of an image are relevant to the evaluation of image quality attributes. For example, flat regions are not relevant for details preservation evaluation. In consequence, in Chapter 3, we design a method that from an annotated set of images according to an image quality attribute, **automatically detect the regions of interest** which are the best suited for the evaluation of this criteria.

- To attain an accurate benchmark of smartphone cameras, it is necessary to conduct a fair comparison by evaluating the cameras on identical content. Synthetic datasets are built with multiple degradations from pristine images, thus providing multiple samples of the same content. However, do not accurately reflect the intricate mixture of defects typically found in images captured by smartphone cameras. Authentic datasets, on the other hand, do encompass diverse images obtained from the web or from camera devices and do include the complex mixture of distortions being sought. However, these datasets have the limitation of presenting only a single content per scene, presenting a challenge for benchmarking purposes that necessitate a precise comparison of quality across a set of contents. As such, there is currently **no available dataset that fully satisfies the requirements for a camera evaluation setup**. To overcome the shortcomings of existing datasets that are either synthetic or do not have a content repetition, **we rigorously create a new dataset** in Chapter 4, tailored to our needs and the elaboration of camera quality assessment evaluation.

- Finally, existing image quality models do not presuppose **the repetition of content across different samples** *ie.* several images depicting the same scene. While this assumption may not be applicable in numerous real-world scenarios, it is a crucial hypothesis to consider in our particular context for the design of more precise models. By acknowledging and accounting for the potential repetition of content, we can strive towards creating models that more accurately reflect the reality of the camera evaluation setup. Our region selection method in Chapter 3 explicitly uses that assumption. Moreover, in Chapter 5, we design a method in that no reference image is available for our use case, but while still retaining our repeated content hypothesis. To this aim,

we propose a novel and original method that leverages several distorted images of the same content in order to build a pseudo-reference.

Throughout this manuscript, we focus specifically on texture, used interchangeably with *details preservation*. It is however most of the time a study case of an image quality attribute evaluation, and the method designed can also be used for other image quality attributes, as seen in Sec. 3.5.

1.3 Manuscript's organization

We start in Chapter 2 with a review of existing industrial standards and methods concerning the evaluation of texture quality and noise levels for cameras. Then we explore existing datasets and explain the most common experimental designs for the annotations of these datasets. Finally, we perform an extensive review of methods aiming to evaluate image quality.

Chapter 3 proposes a region selection method on a chart to evaluate a set of cameras in laboratory conditions. Indeed, evaluating the quality of a set of cameras is typically done by comparing shots of the same visual content, a chart, in a controlled environment. The use of the same chart allows for direct comparisons and easier subjective evaluations. The Modulation Transfer Function (MTF) is a widely used tool for evaluating sharpness and resolution, but it has limitations, including the assumption that the norm of the device transfer function accurately represents texture quality and limitations in evaluating non-linear processing. To address these limitations, we propose DR2S, a deep regression method with region selection, which uses a deep convolutional network to estimate quality scores based on expert human observations. The main novelty is the region selection algorithm which chooses the most appropriate region of the chart for evaluating texture quality. Our findings indicate that, with sufficient training data, learning-based methods outperform the MTF-based method. These results have been published in the International Conference on Pattern Recognition (ICPR)

2020 under the title of DR^2S: "*Deep Regression with Region Selection for Camera Quality Evaluation*" [127].

The last section of Chapter 3 focuses on extending these findings made in the context of texture quality evaluation to the visual noise evaluation. We addressed the challenge of determining the noise level of a camera as perceived by the user, using its processed images. Traditional methods for characterizing camera noise use objective metrics based on charts with uniform patches, but these can result in incorrect characterizations due to the effectiveness of denoising algorithms in uniform areas versus detailed areas. Therefore, we present a method to estimate the perceived noise level on the chart through the use of a deep convolutional network trained on ground-truth quality scores from expert annotators. Our evaluation shows that our approach closely aligns with human evaluations. This chapter's extension was the topic of an article titled "*Automatic Noise Analysis on Still Life Chart*" [9] at the industrial conference London Image Meeting (LIM) in the year 2021.

In Chapter 4 we present a new method for assessing camera quality and a corresponding data collection protocol. The resulting dataset, which is unique in its composition of multiple shots of various natural scenes, allows us to evaluate a camera's ability to capture fine texture details. Reliable image quality annotations are obtained through human comparisons of images depicting the same content in different natural scenes. First, we detail the data collection and annotation protocol. A deep neural network is then trained to predict these human-based quality scores. The chapter concludes with a comprehensive benchmark of existing image quality measures is also performed across various practical scenarios, which provides valuable insights into the advantages and limitations of each method and offers practical recommendations for future research in camera quality assessment. These findings are ready to be submitted.

In Chapter 5, we propose a new no-reference image quality assessment setup and a corresponding method. Unlike existing methods that evaluate each image separately, our method considers multiple images that depict the same content and models them jointly to enhance the accuracy of quality prediction. The idea behind this approach is that multiple distorted

images can provide information to differentiate between image features related to content and quality. To achieve this, we extract the feature representations from each image and use them to create a pseudo-reference, which improves the prediction of quality scores. This work has been presented at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2023 in an article named "*Test your samples jointly: Pseudo-reference for image quality evaluation*" [126].

Chapter 2

Image Quality Assessment Background

In this chapter, we conduct an extensive state-of-the-art review of the literature on image quality assessment. First, considering our focus on cameras, we examine various industrial methods and norms based on test charts shot in a laboratory environment. Second, we discuss the characteristics of publicly available datasets for image quality assessment. Finally, we review existing methods for the two most common settings in the literature: full-reference and no-reference image quality assessment.

2.1 Chart-Based Camera Quality Assessment

A simple model for a camera consists in a linear system that produces an image y as a convolution of the point spread function h and the incoming radiant flux x. In the frequency domain, $Y(f) = H(f)X(f) + N(f)$, where we also consider additive noise N. The modulation transfer function is $\mathrm{MTF}(f) = |H(f)|$ and it is commonly used to characterize an optic acquisition device [13].

Acutance is a single-value metric calculated from the device MTF. To compute this value, a contrast sensitivity function (CSF), modeling the spatial response of the human visual system, is used to weigh the values of the MTF for the different spatial frequencies. The CSF depends on a set of parameters named viewing conditions, which describes how an observer would look at the resulting image. These parameters are usually the printing height

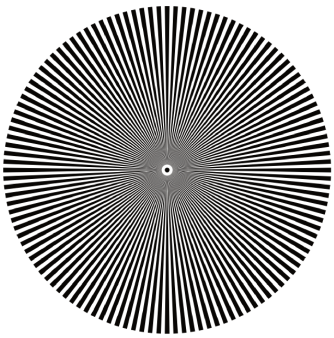


FIGURE 2.1: A Siemens-star pattern to compute the MTF (better viewed with a digital zoom)

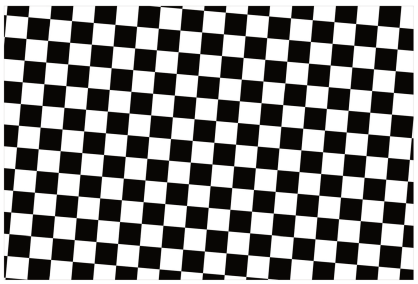


FIGURE 2.2: A grid of slanted-edge patterns for measuring MTF across the field

and the viewing distance. The acutance is defined as the ratio of the integral of the weighted MTF by CSF's integral. More precisely, we have:

$$\text{Acutance} = \frac{\int_0^\infty MTF(v)CSF(v)dv}{\int_0^\infty CSF(v)dv} \tag{2.1}$$

The key assumption in MTF-based methods is that they dispose of the noise-free content to estimate the transfer function. Therefore, these methods are usually used only with synthetic visual charts. To the best of our knowledge, the only work estimating acutance on natural scenes was proposed by Van Zwanenberg et al. [129]. This method uses edge detection in the scene to compute the MTF. Early methods use charts containing a blurred spot or a slanted edge (see Figure 2.2) for this computation. Loebich et al. [82] propose a method using the Siemens-Star (Figure 2.1. Cao et al. propose to use the Dead-Leaves model [88, 45], and introduce an associated method in [19], which is shown to be more appropriate to describe fine detail rendering since its textures are more challenging for devices. This chart is usually referred to as the *Dead-Leaves* chart. In this chart, the reference image is generated with occluding disks with a random center location, radius, and grey-scale value. This pattern is displayed in Figure 2.3

Importantly, digital camera systems present high-frequency noise, which affects the MTF estimation by dominating the signal in the higher frequencies. Consequently, estimating the noise power spectral density (PSD) is key to obtaining an accurate acutance evaluation, and

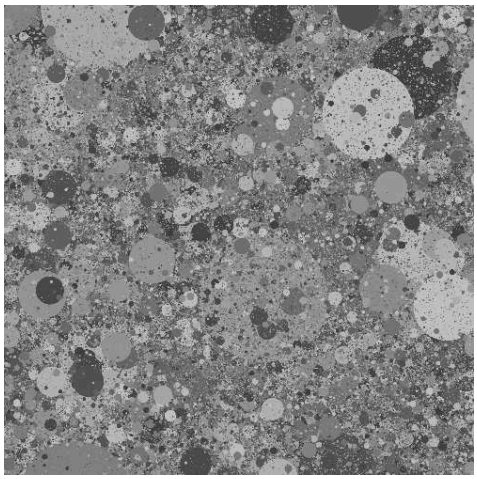


FIGURE 2.3: Example of a Dead-Leaves pattern [19]

this task is not easily performed on the textured region. As a consequence, noise PSD is typically estimated on a uniform patch. However, this approach is hindered by the denoising algorithms integrated into cameras. Not only do these algorithms interfere with the noise PSD estimation but also they behave differently in uniform and textured regions. In this context, Kirk et al. [68] propose to compute the MTF using the cross-power spectral density of the resulting and the reference images. This method assumes an effective registration of the chart. Sumner et al. [119] then modified Kirk's computation in order to make it more robust to slight misregistration.

Concerning noise, its assessment is often done using the signal-to-noise ratio (SNR) as a metric. However, SNR solely measures the overall amount of noise at a given signal level and does not accurately depict how it is perceived by human observers. In response, the visual noise metric has been developed and codified [1, 3, 137]. After a filtering step with specific CSFs for each of the luminance and chrominance channels in the AC_1C_2 [154] color space, the visual noise metric is defined as a weighted sum of the variance in the luminance and chrominance channels in the $CIEL^*a^*b^*$ color space. Parameters of the CSFs and of the final weighted sum are the result of subjective studies. While some standards exist, improvements are regularly proposed [137, 15], leading to variations between different practices. The visual noise metric is usually computed on images of chart with patches of different grey levels as in Figure 2.4.

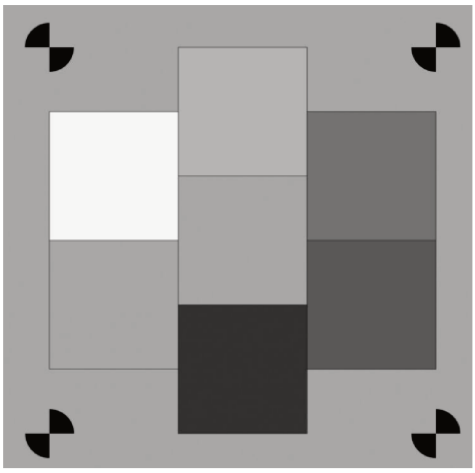


FIGURE 2.4: Chart used to compute the visual noise metric

These methods are widely used and provide valuable hindsight of the camera properties. However, as we explained in the introduction of this manuscript (Chapter 1), we aim to evaluate the image quality of content that is likely to correspond to content photographed by end-users. This motivates our decision to opt for learning-based approaches that can be trained to mimic human preferences. In the next section, we provide a comprehensive review of datasets and methods in this field which we will base ourselves on to adapt and apply to specifically evaluate cameras.

2.2 Datasets

To evaluate image quality assessment methods, datasets comprising images with corresponding ground truths of image quality are required. However, collecting ground-truth quality scores usually involves the conversion of human ratings or comparisons into continuous quality scores, which can be a non-trivial task. Another challenge lies in the acquisition of reliable scores with a defined annotation budget. In the following sections, we review the existing data collection procedures for camera and image quality assessment tasks.

2.2.1 Annotation methods

Different methods of annotation are possible for image quality assessment, with two main families: Mean Opinion Score (*MOS*) and Pairwise Comparison (*PWC*) [145]. The goal of these methods is to convert the quality of an image into a single scalar value.

Rating-based: Mean Opinion Scores

MOS annotation involves asking annotators to individually rate images on a scale from 1 to 5 with integer values. Typically, transformations are applied to the values attributed by the annotator, such as compensating for rating ranges, normalizing for standard deviation in the annotators' ratings, or shifting the value according to the difference between opinion scores between one or several image scores taken as reference. The obtained values are then averaged for each image across annotators.

Pairwise Comparisons

In contrast to MOS annotation, PWC does not ask annotators for a rating of an individual image. Instead, it asks the annotator to choose which of the two images has better quality. Two main technical choices are important for the success of pairwise comparison annotation:

- Pair proposal: To achieve good annotation quality with a limited annotation resource, it is important to propose pairs to the annotators that provide information about the image quality level. For example, images with a large known quality difference are usually not useful to propose as pairs [90].

- Conversion to scalar scores: Converting a set of comparisons to scalar scores is not straightforward, and several possibilities are available. A naive method could involve a championship-like algorithm where the number of wins decides the quality score. In this example, it is required to test every possible pair. Another option is to use a Swiss Tournament system [29, 98], in which images accumulate points for each win and are paired with an image with a similar record, improving the quality of the information provided by each comparison. Finally, it is possible to use an Elo [35] algorithm or one

of the two closely related Thurstone [12] and Bradley-Terry[54] models. These models do not require a particular pair proposal system and allow to use of an effective pair proposal strategy [90]. Used with an efficient pair proposal algorithm, these models appear to be the most efficient [96, 90].

More details on the pairwise comparison annotation method are available in Chapter 4.

When it comes to evaluating the effectiveness of MOS (Mean Opinion Score) ratings compared to pairwise comparisons, Mantiuk et al. [86] have concluded that, when applicable, pairwise comparisons with a forced choice are more accurate and time-efficient.

Annotation conditions

Two different environments can be used for annotation experiments. The first option is a **laboratory** setting [114], which provides precise control over evaluation conditions such as screen calibration, viewing distance, and ambient lighting. Annotators in this setting may or may not be experts in image quality. The second option is **crowdsourcing** [41], also commonly used for image annotation. While these experiments are less expensive to conduct, each participant views the image on their own screen under varying conditions. The trustworthiness of anonymous internet users, as found on platforms such as Amazon Mechanical Turk, is generally considered lower than that of participants in a laboratory experiment [56]. Additionally, these annotators are typically not experts in image quality. In consequence, the design of crowdsourced experiments aims to filter annotation results carefully using test questions with known answers [79, 22] and to offset the lower reliability of individual annotations through a larger quantity of annotations.

2.2.2 Synthetic datasets

Synthetic datasets are generated from a set of natural images considered undistorted or *pristine*, and a set of distortions applied to these pristine images, with different strengths for each distortion.

The **LIVE** [114] image quality dataset from 2006 is still regularly included in methods benchmarks. It comprises of only 5 distortions: jpeg, JPEG2000, white noise, gaussian blur, and fast fading, for a total of 779 images. Due to the low complexity of this dataset, recent methods all obtain high scores on this dataset and in consequence, it is not suited to benchmark recent methods.

A higher number of distortions are present in the **TIDs** [99, 98] datasets with 17 distortions for the 2008 version and 24 for the 2013 version. Distortions correspond to a large number of scenarios, such as image acquisition, compression, watermarking, digital photography, registration, denoising, compression, transmission, inpainting, and reconstruction. Each pristine image is distorted with all the distortions with five levels for each. The resulting 3000 images are rated through a Swiss system.

Compared to the TID2013 [98] dataset, **KADID-10k** [79] is a **considerable size-up with ten thousand rated images**, made from 81 pristine images with 25 distortions and 5 levels. In order to achieve the necessary annotation volume, they conduct a crowdsourced annotation protocol. Many distortions are different from the TID's distortions, mainly including distortion emulating distortions encountered in the wild. The annotation system is rating-based.

The **PIEApp** [101] comprises of around twenty thousand images made from 200 pristine images. They employ **a PWC to annotate samples**. Interestingly, they do not provide scores for images but preference probability for a subset of around eighty thousand pairs.

Used for algorithm evaluation in the challenges of NTIRE workshop of CVPR, the main contribution of **PIPAL** [62] is the addition of the result of **advanced image processing algorithms as distortions**. It includes various algorithms for superresolution and denoising [30, 139, 122], including deep learning-based methods [60, 33] and specifically Generative Adversarial Networks (*GANs*) [109, 152, 47]. In total, it includes around thirty thousand pictures and is rated with an Elo system [35], where rating gains and losses are based on the compared image rating. However, no details on the pair selection process are included by the authors.

2.2.3 Authentic datasets

Even though synthetic datasets include distortions corresponding to acquisition and digital photography use cases, images captured with typical mobile camera devices in real-world scenarios are often affected by a combination of multiple distortions that are not necessarily captured by the synthetic distortions present in existing databases.

The **LIVE-in-the-Wild** [41] dataset was created in 2015 and consists of 1162 images captured from various mobile devices without any artificially introduced distortions. The images were subjected to a large-scale online subjective study, where an average of 175 unique subjects rated each image. To our knowledge, this dataset was the first publicly available to present authentic distortions. Tests of common algorithms on this dataset showed a significant decrease in performance demonstrating the need for such datasets.

However, the LIVE-in-the-Wild [41] is rather small with 1162 images. In 2018, the **KonIQ-10k** [57] dataset is proposed, with **over ten thousand images**. The images are sampled from YFCC100M [121] while ensuring diversity with metrics related to brightness, colorfulness, contrast, sharpness, and content. Each image received at least 120 ratings on a scale from 1 to 5, obtained through crowdsourced online experiments.

With a similar method for image selection and annotation, **PaQ-2-PiQ** [142] proposes in 2019 a dataset four times bigger than KonIQ10k [57]. However, the main originality of this dataset is to also **provide annotation for 120 000 patches** extracted from the images.

Finally, in 2020 Fang *et al.*proposed the **SPAQ** dataset [38], consisting of over 11 thousand pictures taken by 66 mobile devices. Its main quality lies in the annotations: not only the general image quality is annotated, but also five different image attributes as well as scene category labels.

2.2.4 Summary of characteristics

We provide in Table 2.1 a summary of the characteristics of the different publicly available IQA datasets presented.

Dataset	Year	Nature	Image Source	Annotation Method	# of rated images	Environment
LIVE [114]	2006	Synthetic	Degradations	MOS	779	laboratory
TID2013 [98]	2015	Synthetic	Degradations	Swiss-system	3,000	laboratory
PIPAL [62]	2020	Synthetic	Degradations	Pairwise-Comparison	29,000	crowdsourcing
KADID-10k [79]	2018	Synthetic	Degradations	MOS	10,125	crowdsourcing
PIEApp [101]	2018	Synthetic	Degradations	Pairwise-Comparison	20,280	crowdsourcing
LIVE-itW [41]	2015	Authentic	Mobile devices	MOS	1,162	crowdsourcing
PaQ-2-PiQ [142]	2019	Authentic	Internet media	MOS	39, 810	crowdsourcing
KonIQ10K [57]	2018	Authentic	Internet media	MOS	10,073	crowdsourcing
SPAQ [38]	2020	Authentic	Mobile devices	MOS	11,125	laboratory

TABLE 2.1: Comparison of the characteristics of different IQA datasets

2.3 No-Reference Image Quality Assessment

There are two main categories of image quality assessment (IQA) methods: full-reference (FR) and no-reference (NR). FR-IQA algorithms utilize a reference image in order to evaluate the quality of a given test image. This is achieved by calculating the difference between the two images using some form of error metric. As a result, FR-IQA algorithms are widely considered the most accurate and reliable type of IQA method (see Table 2.2 and 2.3). However, they do have the limitation that a reference image must be available in order to perform the assessment. On the other hand, NR-IQA algorithms do not require the availability of a reference image. Instead, they attempt to automatically measure the perceived quality of the test image without human judgment for novel images or the use of a reference image. NR-IQA algorithms can be useful in situations where a reference image is not available or not practical to use.

2.3.1 Distortion Specific Methods

Before 2010, the vast majority of NR-IQA algorithms [16, 87, 135] limited themselves to one or more specific types of distortions, such as blur, blockiness from JPEG compression [135], or ringing arising from JPEG2k compression [87]. As a result, these algorithms have limited application domains. They usually follow this process [115, 123]:

1. Identifying a relevant and discriminative local feature within the image.

2. Using this local feature to model a local distortion metric for the image.

3. Averaging the local distortion metric over the entire image to calculate an overall distortion metric for the image.

4. Using the overall distortion metric to predict an image quality score that aligns with human perception.

From these earlier methods, we will present two: The first one is a very good example of the presented process, while the second one introduces the important concept of Natural Scene Statistics (NSS) As **an example of early methods using handcrafted features**, we succinctly present the method from **Wang *et al.*(2002)**[135]. The aim of this method is to evaluate images distorted by JPEG compression. They only consider the grayscale conversion of the images. The first feature is a measure of blockiness B, defined as the average difference at JPEG block boundaries. A second image attribute to be considered concerning JPEG compression is blurriness. Since blur is difficult to estimate without a reference image, they opt for measuring image "activity". They use two features to characterize the image activity: A, the average absolute difference between in-block image samples, and Z, which corresponds to how often would a line (or a vertical) in the image, considered as a signal, would have its derivative change sign. These four quality metrics are then combined into a single quality metric Q, aligned with human perception: $Q = \alpha + \beta B^{\gamma_1} A^{\gamma_2} Z^{\gamma_3}$ with the 5 parameters fitted to quality scores with a training set.

In these earlier works, the concept of **Natural Scene Statistics** (*NSS*) is reoccurring [16, 113, 92, 108], which assume that perceptual distortions can be measured as deviations of certain statistical properties. More precisely, the image is transformed, commonly with a Discrete Wavelet Transform (*DWT*) or a Discrete-Cosine Transform (*DCT*), and the distributions of the coefficients are considered. This distribution is fitted to a classic distribution, and the fit's parameters are used as features. For example, in **Brandao *et al.*(2008)**[16], propose to use NSS for evaluating the quality of images compressed with JPEG or MPEG. They used an 8×8 blocks DCT transform. For each horizontal/vertical frequency pair, the coefficients

are modeled with a Laplacian probability density function:

$$f_X(x) = \frac{\lambda}{2} \exp(-\lambda|x|) \tag{2.2}$$

with x the coefficient value. Each λ is computed through Maximum-Likelihood. With the help of a training set, they compute the coefficient of a weighted average of λ coefficients for each possible frequency pair.

2.3.2 General Purpose Methods

While the earlier methods were usually crafted for a specific distortion, from 2010 onward, IQA methods have the ambition to evaluate any kind of image impairment.

In 2010, Saad *et al.*[108] proposed the **BLIINDS (2010)** method, which is not designed for specific distortions: **Every distortion can be evaluated with one model** Similarly to [16], they use NSS features from the DCT coefficients for their algorithm. However, they consider the *kurtosis* of the distribution of coefficients. The kurtosis is defined as the fourth standardized moment of a distribution: $\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$. The final quality predictor is crafted through a probabilistic model, using a training set for this purpose.

Another interesting work has been proposed by Moorthy *et al.*[94], named the **DIIVINE (2011)** framework. In this method, they provide **an extensive NSS features** set derived from DWT coefficients. They use in total 88 features, corresponding but not limited to the variance of subband coefficients, the shape parameter of subband coefficients, the shape parameter across subband coefficients, spatial correlation across subbands, and correlations across scales. Their method aims to be multi-distortion, but in reality restricts themselves to five distortions: JPEG2k, JPEG, White Noise, Gaussian Blur, and Fast Fading. With the training set, they train a classification module outputting probability estimates for each distortion. Then, they use different models from the same DWT-based features tailored for each distortion, whose outputs are combined into a single quality score with the probability estimates from the classifier.

Mittal *et al.* [92] with their **BRISQUE (2012)** model was the first research using **NSS features from the spatial domain**, with better results compared to models using features from the frequency domain [108, 94]. Let $I_{i,j}$ be the image pixels' luminance, they study the image with a locally normalized contrast.

$$\hat{I} = \left(\frac{I_{i,j} - \mu(i,j)}{\sigma(i,j) + C} \right)_{i,j} \tag{2.3}$$

with μ and σ Gaussian-weighted mean and standard deviation on a pixel neighboring, and C a constant for numerical stability. Then they fit the distribution of these coefficients and the distribution of the horizontals, vertical, diagonal, and anti-diagonal pairwise product of these coefficients with Symmetric and Asymmetric Generalized Gaussian Distribution. The parameters of the fitted functions are taken as features. For example, the density function of the Generalized Gaussian Distribution is:

$$p(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \, e^{-(|x-\mu|/\alpha)^{\beta}} \tag{2.4}$$

and α_0 and β_0 corresponding to the best fit to one of the aforementioned distributions are part of the feature list.

Rather than fitting features to quality scores with a training set, **NIQE (2012)** [93] use **a set of pristine images NSS to avoid learning**. Using the same handcrafted feature as BRISQUE [92], but instead of training a support vector regression with opinion scores, it utilizes a distance between the distribution of NSS features from patches of the test image and the distribution of the NSS features from a collection of natural pristine images. The distance is directly used as an image quality metric and in consequence, requires no training to obtain this method. **IL-NIQE (2015)** [148] enriches the NIQE model with features related to colors, gradients, and the log-Gabor responses of the image.

Ye *et al.* [140] took a first step toward **data-driven features** with their **CORNIA (2012)** model. From randomly selected image patches on the image database, normalized and whitened, which means performing component analysis over image channels in order to

decorrelate them, they use the k-mean clustering algorithm to select 20 000 centroids as descriptors. Note that image quality scores are not taken into consideration during the feature elaboration. The method can in consequence be labeled as unsupervised. A Support Vector Regression (SVR) is then trained using maximum dot-product values over the patches of each feature in the codebook as inputs. In a later 2013 article [141], authors reduce the number of features used to 100 through a supervised learning procedure, using quality scores for the feature selection.

2.3.3 Convolutional Neural Networks-based methods

Following the success of convolutional neural networks in image classification [71], these were since tested for the image quality assessment task.

In **2014**, **CNN-IQA** [65] has been developed by Kang *et al.*and **used for the first time a Convolutional Neural Network (CNN) for general-purpose IQA**. The goal is to predict the Mean Opinion Score (MOS) of the images. The CNN takes as input 32×32 random patches whose ground-truth quality scores are set as the same as the source image score. The proposed architecture is rather small: There is a single convolutional layer comprising 50 kernels. At test time, the prediction for an image is the average of patch-wise predictions. This work achieves the 2014 state of the art in both the LIVE dataset and TID2008 at the time of publication. This success sparked further research into using convolutional networks for IQA.

Later works propose more complex architecture as WaDIQaM-NR [14] by Bosse *et al.*. **WaDIQaM (2017)** [14] architecture includes 10 convolutional layers with a max-pooling layer for every other convolutional layer for feature extraction and 2 fully connected layers for regression. Similarly to CNN-IQA [65], the inputs are 32×32 random patches, even though the architecture is more complicated. However, an interesting feature of WaDIQaM is the patch-weighting estimator to provide **an estimation of the importance of each zone**: Indeed the quality of an image as perceived in a local region may not necessarily be indicative

of the overall quality of the image. This can be due to various factors such as spatially non-uniformly distributed distortion or saliency effects. To address this issue, they propose a patch-weighting estimator: in parallel to the regression module, with the same feature as inputs, a second regression module outputs weights to be applied to the patch. Weights from all patches are normalized to have a unitary sum and used to compute a weighted average of individual patch scores. Nonetheless, their experiments from the paper show that the gain in performance due to the patch-weighting estimator is rather small.

Other approaches rather choose to focus on the target quality scores modelization and adopt **a probabilistic representation of scores**, modeling the distribution of the quality scores of each image. An interesting attempt has been done by Zeng *et al.*[144] with the **PQR (2018)** method. Instead of performing a regression to a scalar quality score, they divide the quality range into small score bins or quality anchors. To this aim, they transform the ground truth from the training dataset: Instead of a single quality score, the ground truth is modeled through a probabilistic formulation. The probability at each quality anchor is derived from considering a Gaussian density function, centered at the real ground truth. Since their network predicts a vector and not a single quality score as desired, they further train a linear support vector regression to map the vector of probabilistic representation to quality scores.

Similarly, Talebi *et al.*[120] also uses distributions as ground truths for image quality instead of a single score representation in the **NIMA (2017)** method. However, to the contrary of PQR[144], they have not adopted a probabilistic formulation, but instead used annotations from different annotators into different quality levels directly, **using the actual distribution of annotations**. Considering $\mathbf{p} = [p_1, \ldots, p_N]$ the rating's distribution across N score buckets, they train the network output $\hat{\mathbf{p}}$ to match $\mathbf{p}$, using a distribution distance as a loss function. More precisely, they used the Earth's Mover Distance (EMD):

$$\mathrm{EMD}(\mathbf{p}, \hat{\mathbf{p}}) = \left(\frac{1}{N} \sum_{k=1}^{N} \left| \mathrm{CDF}_{\mathbf{p}}(k) - \mathrm{CDF}_{\hat{\mathbf{p}}}(k) \right|^r \right)^{1/r} \tag{2.5}$$

with $r = 2$. If a single scalar is needed for a quality score prediction, they propose to simply use the mean defined as $\mu = \sum_{i=1}^{N} s_i \times \widehat{p}_i$ with s the score buckets values.

The premise of the work from Zhang *et al.*[153] is the observation that different approaches perform differently regarding the difference in the nature of the data. They claim that while fine-tuning directly a model trained on ImageNet [107] with an IQA dataset images performs well for authentic distortions but the performances are not outstanding on synthetic distortions dataset. Conversely, patch-based training fails to handle authentic distortions properly due to their non-homogeneity. With their **DBCNN (2019)** architecture [153], they aim to **handle better both authentic and synthetic distortions**. They propose another *no-reference* method that is composed of two different convolutional models: the first one is trained on synthetic distortion classification and the second is a pre-trained VGG [116] on ImageNet. This pre-trained model is introduced to better represent the perceptual features of natural images. The two sets of features are merged into a single representation for a final quality prediction with a bilinear pooling: Considering representations of the first network $\mathbf{Y}_1$ of size $(h \cdot w) \times d_1$ and $\mathbf{Y}_2$ of size $(h \cdot w) \times d_2$ for the second one, they used the merged representation $\mathbf{B} = \mathbf{Y}_1^T \mathbf{Y}_2$ as input of the quality estimator regression, which is the only part of the network trained on the image quality dataset.

The authors of **HyperIQA (2020)** [118], Su *et al.*, observed that currently, existing NR-IQA models predictions do not take into consideration the vast diversity of images. Once trained, they used the same predictor to transform features into a quality score. It implies that the same sort of image quality features is needed for predicting diverse images. They argue that this is not the case in practice, and use the very telling example from [77] of an image depicting a clear blue sky: while humans may regard this image as high-quality, models will tend to evaluate it as a blurry image. They conclude that it is important to **understand the context of the evaluation**. In consequence, different rules need to be applied for predicting the image quality depending on the content. To this aim, they use a hyper-network, taking as inputs only the semantic features (corresponding to the last convolutional layers of the network). They train this hyper-network to output adapted weights for the fully convolutional layers.

Zhu *et al.*[157] with the **MetaIQA (2020)** proposed to **apply meta-learning techniques** to the problem of image quality assessment, to better **adapt to unknown distortion**. From a training set consisting of various distortions, they seek to provide initialization weights from which they will be able to train the network highly efficiently with few examples and a few gradient updates on a new distortion. The algorithm used for this task is a first-order meta-learning algorithm, inspired by FOMAML [39].

Another popular technique in computer vision is the contrastive learning method. Contrastive learning refers to the technique of leveraging a large amount of unlabeled data and achieve to use this data to pre-train a model in a self-supervised manner, learning useful representations for the task at hand. **CONTRIQUE (2021)** [84], proposed by Madhusudana *et al.*applied these **Contrastive pretraining** techniques to the image quality assessment problem. For the contrastive pre-training, they use the unlabeled KADIS-700k dataset, which consists of pristine images and their degraded version with different distortions, and intensity of distortion. Considering degradation $\times$ intensity couples, they consider each unique couple as a different distortion class. The contrastive objective consists of decreasing the representation distance in the feature space between pairs of the same class. To this aim, they use the NT-Xent loss [23] for positive pair couples:

$$\mathcal{L}_{\mathrm{NT-Xent}} = -\frac{1}{N}\sum_{i,j}\log\frac{\exp\left(\mathrm{sim}\left(z_i, z_j\right)/\tau\right)}{\sum_{k=1}^{2N}\mathbb{1}_{[k\neq i]}\exp\left(\mathrm{sim}\left(z_i, z_k\right)/\tau\right)} \tag{2.6}$$

with sim the cosine-similarity, N the number of positives pairs and τ a temperature parameter. However, this first contrastive objective only considers synthetic distortions. In consequence, for authentic distortions, they consider a mix of various datasets, not necessarily designed for image quality, and consider pairs to be positive if they are quality-preserving transformations of the same original image. The contrastive pretraining uses at the same time both sets of images. For the actual training on quality scores of images from an image quality dataset, they use a ridge regression [55] to map these learned features to quality scores.

2.3.4 Vision-Transformers for IQA

While previous approaches were based on convolutional networks, **MUSIQ (2020)** [66] employs the recent **vision-transformer architecture**[34]. Their improvement over the standard ViT architecture consists in ingesting patches interpolated at different scales. The location token is computed from a look-up table common to every scale.

Another work from Golestaneh *et al.*[44] proposed the novel **TReS (2022)** architecture to tackle the no-reference image quality problem **employing both a CNN and a Transformer encoder**. Starting with a CNN-feature extractor at 4 different stages of the network, they feed these features to a transformer encoder. Interestingly, they also use a self-consistency loss, ensuring different augmented versions of an image have similar representation both after the convolutional part and the transformer encoder.

2.3.5 Performances on common datasets

In Table 2.2, we display the performances of the no-reference methods on four of the presented datasets. Most metrics were taken from [84] and [118], since they used the same evaluation protocol and had extensive state-of-the-art comparisons. In other cases, performances come from the original papers. While for the LIVE [114] dataset, early methods such as BRISQUE [92] and IL-NIQE [148] achieve comparable performances to learning-based methods, results on TID2013 [98] and KonIQ-10k[57] show more contrasted performances. On the LIVE-in-the-Wild [41] dataset, results of non-deep learning-based methods this contrast of performances is even more apparent.

Method	LIVE	LIVE-ITW	TID2013	KonIQ-10k
BLIINDS [108]	0.912	0.405	0.536	0.585
DIIVINE [94] []	0.925	0.508	0.549	0.589
BRISQUE[92]	0.939	0.608	0.604	0.665
NIQE [93]	0.907	0.455	0.315	0.531
IL-NIQE [148]	0.946	0.432	-	0.507
CORNIA [140]	0.947	0.629	0.678	0.780
CNN-IQA [65]	0.956	-	-	0.572
WaDIQaM [14]	0.960	0.682	0.835	0.804
PQR [144]	0.965	0.857	0.740	0.880
NIMA [120]	-	0.637	0.750	-
DBCNN [153]	0.968	0.851	0.816	0.875
HyperIQA [118]	0.962	0.859	0.840	0.906
CONTRIQUE [84]	0.960	0.845	0.843	0.894
TReS [44]	0.969	0.846	0.883	0.915

TABLE 2.2: Comparison of state-of-the-art NR-IQA algorithms. The *SROCC* is reported.

2.4 Full-Reference Image Quality Assessment

2.4.1 Traditional Methods

The mean squared error (*MSE*) and its associated metric, the peak signal-to-noise ratio (*PSNR*), are the most basic full-reference image quality metrics. These metrics are favored due to their ease of calculation, clear physical interpretations, and mathematical simplicity in the context of optimization. However, they have a limited correlation with perceived visual quality and therefore may not accurately reflect the human experience of image quality.

Traditional methods share common characteristics [134, 149, 147]: they select a set of features that can be computed separately for a sub-window of the image. Secondly, for a feature F evaluated for a sub-window x, they compute the feature for the same sub-window of two images, F_1 and F_2. The similarity concerning this feature is calculated as follows:

$$S_{F(x)} = \frac{2F_1(x) \cdot F_2(x) + c}{F_1(x)^2 + F_2(x)^2 + c} \tag{2.7}$$

with c a constant to be determined. This similarity can be extended to the case of multiple features F by considering their product. The importance of each feature can be adjusted by introducing weighting parameters as exponents [134]. The results for sliding windows are then pooled across the whole image to create a single value of similarity.

The **Structural SIMilarity Index (SSIM, 2004)** [134] is intended to measure the similarity between two images **by comparing their structural content**. Specifically, it consists of three separate indices related to luminance, contrast, and structural information, all of which are known to be important for the Human Visual System (*HVS*). The luminance index is related to the mean intensity of pixels in both images and is computed on a small window, typically 11×11. The contrast index is related to the standard deviation of pixel luminance over both windows, while the structural index uses the covariance of pixel luminance. To obtain a metric that compares the two full images, the metric is averaged over sliding windows. The **Multiscale-SSIM (MS-SSIM)** [136] improves upon the SSIM by computing the SSIM at different scales and weighting them to a single metric. The weights are determined by the authors through a subjective study.

Using a similar approach to the SSIM [134], the **Feature SIMilarity index (FSIM, 2011)** [149] is a composite index of two features known to be important for the HVS. First, they use the **phase congruency** [70] feature, which relates to the significance of a local structure. Specifically, it is defined as how aligned the phases of the Fourier decomposition of the image are. Second, they use the gradient magnitude as their second feature, which is known to be important for humans as it carries the contrast information, and the phase congruency is contrast invariant. The FSIM achieves noticeably better performance than the SSIM.

Zhang *et al.*[147] proposes to use the concept of **saliency maps** [58, 17, 112, 146] to estimate the importance of different regions of an image for the HVS. In their method, the purpose of Visual Saliency is twofold: first, it provides a feature map that describes the local quality of the image; second, it serves as a weighting function that determines the relative importance of a local region in the quality score pooling process. In their **Visual Saliency-induced Index (VSI, 2014)**, they use the SDSP method [146] to compute the saliency map,

and combine it with a gradient magnitude feature to account for contrast information, similar to the FSIM method [149].

Another approach uses Haar wavelets [49] to build their features. Using horizontal and vertical Haar wavelets filters at three scales, the **HaarPsi (2018)** [104] measure does not use any other feature. In addition to being **more performant** than the VSI [147] and FSIM [149] on most datasets, it is three to eight times **faster to compute** than any of those metrics.

2.4.2 Learning-based Methods

Similarly to NR-IQA, the best-performing methods in full-reference IQA gradually shifted from handcrafted metrics based on the human visual system to learning-based methods. FR-IQA method based on deep learning is generally based on the following steps:

- Process the reference image and the image to be evaluated with the same backbone convolutional neural network to produce a set of features for each of the images. This architecture is commonly described as a siamese network [69].
- Aggregate the two sets of feature to either feed the resulting features to a regressor or directly uses the feature to compute a distance between the two images.

While the backbone networks used followed the evolution of backbones used in computer vision, we will focus on the differences in the aggregation of the two sets of features.

The full-reference version of **WaDIQaM (2017)** [14] (See Section 2.3), concatenates the features from both images and the difference between these features. They compared it with the concatenation only and the difference only and concluded that this aggregation provides better performances.

The **LPIPS (2018)** [150] is an influential full-reference method often used either as a metric [10] or as a loss in tasks such as image synthesis [36, 106, 91], super-resolution [63] or view rendering [102, 132]. Using a pre-trained backbone, typically a VGG-16 [116] on ImageNet, they show that learning only a linear regressor is sufficient for good performance.

Concerning the feature aggregation, they normalize the activations in the channel dimension, scale each channel by a learned weight, and then compute the Euclidean distance between the reference and degraded image. This process is done at several scales across the network and the distances are then averaged. On aggregation schemes, while **DualCNN (2020)** [131] concatenates features from min-, average-, and max-pooling, Varga [130] (2020) obtains good results with the use of traditional **full-reference metrics directly applied on a CNN's features maps**, such as the SSIM [134] or the HaarPsi [104] metrics. Authors of **DISTS (2020)** [32] propose an interesting scheme for feature aggregation. To combine two feature maps from the reference and the image to be evaluated, they compute both the structure and luminance part of the SSIM [134] for each channel. These obtained features are then combined linearly with learnable weights for all the channels for the outputs of 5 convolution layers corresponding to different resolutions.

Concerning backbones, recently proposed methods from the NTIRE challenge [46] use **hybrid architecture featuring both convolutional networks and transformer encoders**. Lao *et al.*[72] (2022) uses both a CNN and a Vision-Transformer [34] to process the two images and then combine features to feed another CNN. Cheon *et al.*[26] (2021) uses a transformer encoder to process features from a CNN. Several other works base their method on this architecture [28, 27].

2.4.3 Performances on common datasets

In Table 2.3, we display the performances of the full-reference methods on four of the presented datasets. Some metrics were taken from [84] which performed extensive state-of-the-art comparisons. In other cases, performances come from the original papers. Even more so than the no-reference results (Table 2.2), results on the LIVE [114] dataset fail to discriminate methods, while clear performances gains are observed on TID2013 [98] and KADID-10k [79] over the years of methods improvement.

Method	LIVE	TID2013	KADID-10k
PSNR	0.881	0.643	0.677
SSIM [134]	0.948	0.637	0.724
MS-SSIM [136]	0.951	0.787	0.802
FSIM [149]	0.964	0.852	0.854
VSI [147]	0.951	0.902	0.880
HaarPsi [104]	-	0.873	0.885
LPIPS [150]	0.932	0.673	0.721
PieAPP [101]	0.915	0.877	0.869
DISTS [32]	0.953	0.942	-
CONTRIQUE-FR [84]	0.966	0.909	0.946

TABLE 2.3: Comparison of state-of-the-art FR-IQA algorithms. The *SROCC* is reported.

Chapter 3

Fixed Content Discriminant Region Selection

3.1 Introduction

A typical way to evaluate the quality of a set of cameras consists of comparing shots of the same visual content in a controlled environment. The common visual content is usually referred to as a *chart*. Chapter 2 displays several charts and details the associated computations. The motivation for using the same chart when comparing different cameras is twofold. First, it facilitates the direct comparison of different cameras. In particular, when it comes to subjective evaluation, humans can more easily provide pairwise preferences than an absolute quality score. Second, when the common noise-free chart is known, this reference can be explicitly included in the quality measurement process.

In this context, the Modulation Transfer Function (MTF) is a widely used tool for evaluating the perceived sharpness and resolution [13], which are essential dimensions of texture quality. First, MTF-based methods suffer from important drawbacks. MTF-based methods are originally designed for conventional optic systems that can be modeled as linear. Consequently, non-linear processing in the camera processing pipeline, such as multi-image fusion or deep learning-based image enhancement, may lead to inaccurate quality evaluation [129].

Second, these methods assume that the norm of the device transfer function is a reliable measure of texture quality. This assumption fails to account for many nuances of human perception. Some recent works have shown that the magnitude of image transformations does not always coincide with the perceived impact of the transformations [151]. Consequently, in this chapter 3 we advocate that human judgment should more explicitly be included in the texture quality measurement process.

As a consequence, we propose DR^2S, a Deep Regression method with Region Selection. Our contributions are threefold. First, we formulate the problem of assessing the texture quality of a device as a regression problem and we propose to use a deep convolutional network (ConvNet) for estimating quality scores. We aim to obtain a score that would be close to a subjective quality judgment: to this end, we use annotations provided by expert human observers as ground truth at training time. Second, we propose an algorithm to identify the regions in a chart that are better suited to measure perceptual quality. Finally, we perform an extensive evaluation study that shows that our learning-based approach performs better than existing methods for texture quality evaluation and that our region selection algorithm leads to further improvement in texture quality measurement.

3.2 Related works

Existing methods can be classified into two main categories: MTF-based and learning-based methods. An extensive review of the MTF-based methods is available in Chapter 2. We summarize here the methods included in our experimental evaluation. The acutance is a single value metric defined as the weighted integral of the MTF with a contrast sensitivity function modeling the spatial response of the human visual system. Cao et al. propose a method [19] using the Dead-Leaves model [88, 45] which is more appropriate than a simple slanted-edge chart to describe fine detail rendering as it is more challenging for devices. Estimating the noise power spectral density (PSD) is key to obtaining an accurate acutance evaluation, as high-frequency should not be mistaken for fine details. This task is not easily performed on the textured region, as a consequence, noise PSD is estimated on a uniform patch. It

is important to note that only the PSD of the reference image and not the reference image itself is needed for the acutance computation. For this reason, this method is referred to as Reduced-Reference (RR) acutance in the rest of this chapter. However, denoising algorithms integrated into cameras behave differently in uniform and textured regions. In this context, Kirk et al. [68] propose to compute the MTF using the cross-power spectral density of the resulting and the reference images, assuming effective registration of the chart. Sumner et al. [119] then modified Kirk's method to improve slight misregistration robustness. Since this method takes full advantage of the reference image, it is referred to as Full-Reference (FR) acutance in the rest of this chapter.

In conclusion, state-of-the-art techniques typically allow us to obtain a good estimation of devices' MTF and then of the acutance. However, it has been shown that acutance itself does not always reflect very well the human quality judgment [85]. This observation calls for learning-based methods that aim at reproducing the score of human experts evaluating the images.

Concerning learning-based methods exposed in Chapter 2, they tackle a slightly different problem than ours. They are designed to evaluate image quality attributes for any input image while we address the problem of evaluating devices using a known common chart.

3.3 Method

3.3.1 DR^{2}S overview

In this section, we detail the proposed method for estimating texture quality. We formulate this task as a regression problem. We assume that we dispose of a training dataset composed of N color images $(X_1, \ldots, X_N)$ of dimensions $H \times W$ with the corresponding texture quality scores $(Y_1, \ldots, Y_N) \in \mathbb{R}^N$. Since we are interested in estimating the camera quality, each score Y_n corresponds to a quality score for one device at a specific lighting condition. Note that, several training images can be taken with the same device. Importantly, we aim

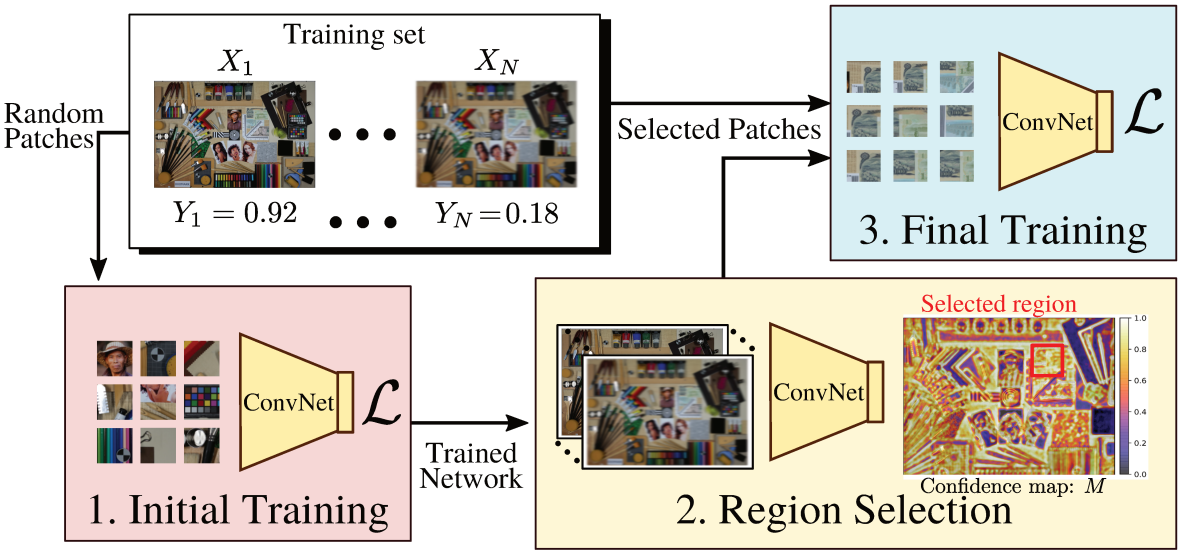


FIGURE 3.1: Pipeline overview of DR^{2}S: Our method is divided into three main stages. First, we perform naive training using patches randomly cropped over the chart. Second, we compute a map that indicates discriminant regions. Third, we perform a final training using only a selected region.

at computing quality scores that coincide with human judgment, the ground-truth texture quality scores are provided by human annotators (See Sec.3.4.1 for more details). We aim at training a ConvNet $\phi(\cdot, \theta)$ with parameters θ. We consider that these images depict the same chart and are taken with different cameras and different lighting conditions.

Our method is based on the observation that all image regions are not equally suited to predict the overall image quality. For example, regions with uniform textures will be rendered similarly by any device independently of its quality. Conversely, other regions with rich and fine texture details are much more discriminating since they will be differently captured by different devices. Based on this observation, we propose the algorithm illustrated in Fig. 5.1. The method is divided into three stages: First, we train a first neural network ignoring this problem of unsuited regions. Second, we employ this network for estimating which image regions are the most suitable for measuring image quality. Finally, the network is re-trained using only selected regions. In the following, we provide the details of the three stages of our algorithm.

3.3.2 Initial training

The goal of this first stage is to train a neural network that can be used to identify relevant image regions. To this end, we propose to train a network to regress the quality score from an input image patch. This initial network is later used in the second stage of our pipeline to identify discriminant regions (See Sec.3.3.3). We train this deep convNet on random crops extracted from the training images $\{X_1, \ldots, X_N\}$. These crops are randomly selected across the images with a uniform distribution. In all our experiments, we use the widely used Resnet-50 network pre-trained on ImageNet [31] where the final classification layer is replaced by a linear fully connected layer. Since some patches are not discriminant, this initial training suffers from instability and optimization issues. Consequently, we use the Huber loss that reports better performance in the presence of noisy samples [20]:

$$\mathcal{L}(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{for } |y - \hat{y}| \leq \delta \\ \delta |y - \hat{y}| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases} \tag{3.1}$$

where y and $\hat{y}$ denote the annotated and predicted scores, and $\delta \in \mathbb{R}^+$ is a threshold. At the end of this stage, we obtain a network that estimates the quality of a device from an input patch.

3.3.3 Region selection

In the second stage of our pipeline, we use our previously trained network to predict image regions that are most discriminating for quality measurement. We produce a map $M \in [0, 1]^{H \times W}$ that indicates the relevance of each location of the chart to estimate texture quality. This map will allow us to select a suitable region to train our convNet in the last stage of our pipeline.

In order to estimate a single map M for all the training images, we first register all the training images. We employ the following algorithm to align the images on the image with the highest resolution (47 MP). First, we detect points of interest. Then, we extract

local AKAZE descriptors [6] and, finally, we estimate a homography for every image. Image warping is implemented using bicubic sampling. Note that, while the map computation requires this warping alignment step that may affect the performance, training and prediction can be performed on the original images. We now assume that the training images $\{X_1, \ldots, X_N\}$ are registered.

To estimate the map M, we propose to use the network ϕ trained in the first stage. Let $\Psi \in \mathbb{R}^{H',W',C}$ be the feature tensor outputted by the backbone network for a given input image. In our case, since we employ a ResNet-50 network, Ψ corresponds to the tensor before the Global Average Pooling (GAP) layer. Since Resnet-50 is a fully convolutional network, the dimension H' and W' depends on the input image dimensions, while the number of channels is fixed (i.e., $C = 2048$). Let $A \in \mathbb{R}^C$ and $b \in \mathbb{R}$ be the trained parameters of the final regression layer obtained in the first stage. The network prediction is given by:

$$\hat{y} = \sigma\left(A^\top \cdot \text{GAP}(\Psi) + b\right) \tag{3.2}$$

where σ denotes the sigmoid function. While the network returns one single output scalar per input image, we want to obtain one value per pixel location. In order to adapt the class activation map framework [156] to our regression setting, we propose to compute a score for every feature map location $(h, w) \in [1 : H'] \times [1 : W']$:

$$S_{h,w} = \sigma\left(A^\top \cdot \Psi[h, w]) + b\right) \tag{3.3}$$

where $\Psi[h, w] \in \mathbb{R}^C$ denotes the feature vector at the location (h, w). Note that, the resulting map $S = (S_{h,w})_{(h,w)\in[1:H']\times[1:W']}$ has a dimension $H' \times W'$ that is different from the initial input image size $H \times W$. This size difference depends on the network architecture. In the case of Resnet-50, we obtain a ratio 32 between the input and the feature map dimensions. Therefore, we resize the score map S to the dimension $H \times W$ using bicubic sampling. This procedure is applied to every image of the training set. Thus, we obtain the set of score maps $\{\tilde{S}_1, \ldots, \tilde{S}_N\}$

We propose to define the confidence score map M as the location-wise variance of the score maps $\tilde{S}_n$ over the whole training set. The motivation for this choice is that, discriminating regions have a higher variance than non-discriminating ones. Indeed, we observe that the scores produced by the ConvNet $\phi(\cdot, \theta)$ over non-discriminative regions tend to have small variance. Conversely, on discriminating patches, the networks predict values with a wide range leading to high variance.

3.3.4 Final training and prediction

In the last stage of our pipeline, we select the chart region with the highest confidence score value in M. In our preliminary experiments, we observed that using a region width approximately six times larger than the network input size leads to good performance. In our case, we use a square region of 1200×1200 pixels. In this region, we select random patches that are used as a training set. We re-train the network ϕ, starting again from ImageNet pre-training weights.

At test time, assuming an image with an unknown quality score, we extract patches in the selected region. The final score is given by the average over the different patches.

3.4 Experiments

In this section, we perform a thorough experimental evaluation of the proposed pipeline. We implemented our method using Tensorflow and Keras. When training the ConvNets (stages 1 and 3), we employ Adam optimizer following [74], with a starting learning rate of $3 \cdot 10^{-3}$ with a decay of 0.1 every 40 epochs for a total of 120 epochs. In our model, we assume that all the images have the same resolution. The reason for this choice is that we want the image details to be analyzed at the same scale, as a human observer would do. In practice, the resolution depends on the device. Therefore, we preprocess all the training images resizing them to the highest resolution of the dataset using bicubic upsampling. This solution is preferred to downsampling to a common lower resolution since texture quality

is not invariant to downsampling. In addition, due to possible lens shading, we remove the sides of the images.3.1: Ablation Study: we measure the impact of region selection com-

3.4.1 Datasets

Charts and devices As there is no well-established reference dataset for our problem, we collected annotated data using three different charts.

- **Still-Life**: First, we use the chart displayed in Fig. 3.3a. This dataset is referred to as *Still-Life*. The chart is designed to evaluate several image quality attributes and to present diverse content: Vivid colors for color rendering, fine details, uniform zones, portraits as well as resolution lines, and a low-quality Dead-Leaves version. Images are acquired using 140 different smartphones and cameras from different brands commonly available in the consumer market. In Fig. 3.2, we provide an example patch captured using three different cameras. The left image corresponds to a high-quality device while the two others are obtained with low-quality devices. It illustrates the nature of distortions that appear in this dataset with different intensities. To obtain a larger database and predictions robust to lighting conditions, we shoot the chart using five different lighting conditions: 5 lux tungsten, 20 lux tungsten, 100 lux tungsten, 300 lux TL84, and 1000 lux D65. Note that the process is repeated for every device.

- **Gray-DL**: Second, we employ the dead-leave chart proposed in [19]. As mentioned in Sec 3.2, this chart depicts gray-scale circles with random radii and locations. In all our experiments, we refer to this dataset as *Gray-DL*. We use the same five lighting conditions and devices as for the *Still-Life* chart.

- **Color-DL**: Finally, we complete our experiment using the dead-leaves chart proposed in [119]. In opposition to *Gray-DL* this chart is colored (an image can be found at [2]). For this chart, we employed a limited number of devices. More precisely, we employ only 14 devices with the same five lighting conditions as for the other charts. The low number of devices is especially challenging for learning-based approaches.

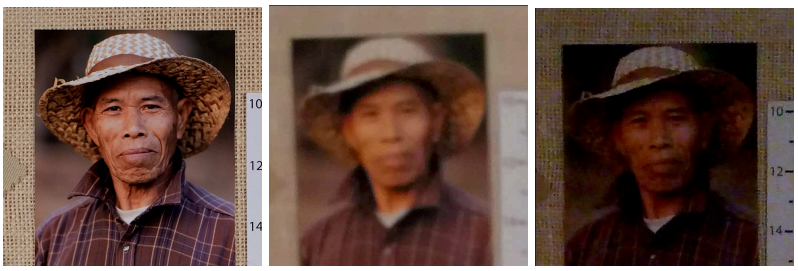
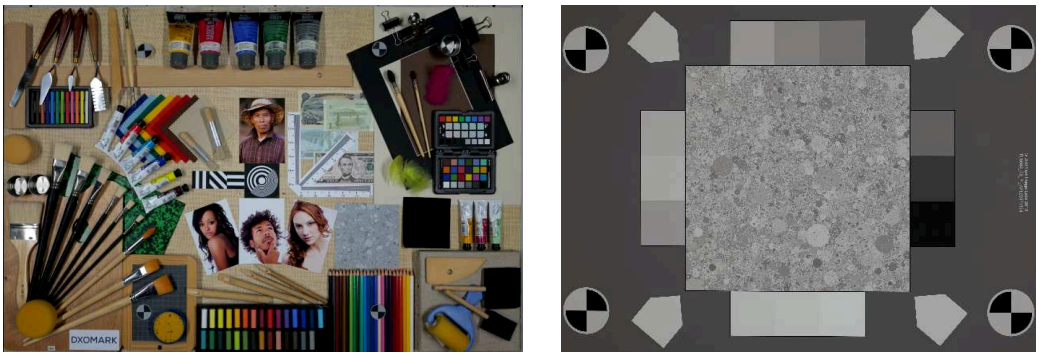


FIGURE 3.2: Patches of high (left) and low-quality (center and right) images from our *Still-Life* dataset



(A) *Still-Life* Chart (B) *Gray-DL* Chart

FIGURE 3.3: Still-Life Chart used in our experiments. The *Still-Life* chart contains many diverse objects with varying colors and textures while the *Gray-DL* chart depicts random gray-scale circles.

Annotations In order to train and evaluate the different methods we need to provide ground-truth annotation for each device. Note that, the annotation must be provided for each pair of device and lighting conditions. To obtain quality annotations that are a reliable proxy of the perceived visual quality, annotations are provided by human experts. Images to be evaluated were inserted among a fixed set of 42 references, and very high-quality prints were provided to help them judge the authenticity of the details. Annotators were asked to compare the images using the same field of view for every image, using calibrated high-quality monitors where images are displayed without applying any down-sampling but with a possible digital zoom for the lower resolution image. Each position among the set of references is assigned a score between 0 and 1. In the case of the Dead-Leaves charts, since the charts are unnatural images, human perceptual annotation is problematic. Therefore, we chose to use the annotations obtained on the *Still-Life* also for the dead-leaves charts, rather than annotating

the images. The *Still-Life* chart contains diverse textures similar to what real images would contain. In this way, we obtain a subjective device evaluation in a setting more similar to real-life scenarios.

3.4.2 Metrics

In our problem, relying on standard classification or regression metrics is not straightforward. Indeed, MTF-based methods predict a quality score that is not directly comparable to the score provided by human annotators. A straightforward alternative could consist in computing the correlation between the predictions and the annotation. However, the underlying assumption that the predictions of each method correlate linearly with our annotations may not hold and bias the evaluation. Therefore, we decided to rely on two distinct metrics based on the correlation of the rank-order. First, we adopt the Spearman Rank-Order Correlation Coefficient (*SROCC*) defined as the linear correlation coefficient of the ranks of predictions and annotations. Second, we report the Kendall Rank-Order Correlation Coefficient (*KROCC*) defined by the difference between concordant and discordant pairs divided by the number of possible pairs. The key advantage of this second metric lies in its robustness to outliers.

For all visual charts, the dataset is split into training and test sets as follows. First, among the devices we use in our experiments, several are produced by the same brand. So, to avoid bias between training and test, we impose no brand overlap between training and test sets. Second, as a consequence of such constrain, a limited number of brands may appear in the test set. To avoid evaluation biases towards specific brands, we use a k-fold cross-validation with $k = 16$.

In order to measure the impact of the number of devices on performance, we perform experiments with a variable number of devices. For all the experiments on the *Gray-DL* and *Still-life* charts, we report the results obtained using subsets of size 20, 60, 100, and 140 devices. For a given number of devices, each experience is performed over the same device

Number of devices	20	60	100	140	20	60	100	140
			SROCC				**KROCC**	
Random Patch	0.626	0.818	0.784	0.806	0.433	0.617	0.588	0.613
Random Region	0.795	0.863	0.866	0.879	0.606	0.680	0.682	0.700
Selected Region (Full model)	**0.830**	**0.912**	**0.890**	**0.900**	**0.638**	**0.740**	**0.716**	**0.728**

TABLE 3.1: Ablation Study: we measure the impact of region selection by comparing three baseline models on the *Still-life* chart. *SROCC* and *KROCC* metrics are reported.

set. Note that, for every method, the complete pipeline is repeated independently for every subset.

3.4.3 Ablation study

In order to experimentally justify our proposed method, we compare three different versions of our model:

- *Random Patch*: In this approach, random patches are selected from the whole chart at both training and testing time.
- *Random Region*: We then restrict the random patch extraction to a single zone, chosen randomly. We report the average over five random regions.
- *Selected Region*: In this model, we employ our full pipeline as described in Sec. 3.3. In particular, training and test are performed using the selected region.

In these three models, we employ a ResNet-50 backbone trained using the same optimization hyper-parameters.

The results obtained on the *Still-life* chart are reported in Table.3.1. First, when using the *random patches* variant, the model trained on 20 devices performs poorly both in terms of *SROCC* and *KROCC* compared to other variants. In this case, we see that it is required to dispose of at least 60 devices to get satisfying performances. Second, we observe that restricting random patches extraction to a region randomly selected leads to better performance than if

we do not restrict to this region. The gain is visible for every number of devices and for both metrics. It may be explained that the decreased diversity in content leads to a ConvNet that is specialized in a specific region of the chart. In other words, the benefit of a more restrained input diversity is larger than the benefit of a larger and more diverse training set. Finally, our full model reaches the best performance for both metrics and for every number of devices. This better performance independently of the training sub-set demonstrates the robustness of the proposed method. Interestingly, we obtain performances with 20 devices similar to the performance of the *Random Patch* model with 140 devices. Overall, this ablation study illustrates the benefit of selecting specific regions for texture quality measurement.

3.4.4 Qualitative analysis of our region selection

In order to further study the outcome of our region selection algorithm, we display the resulting map (Fig. 3.4) of relevant zones. We observe in Figure 3.4 that uniform regions are

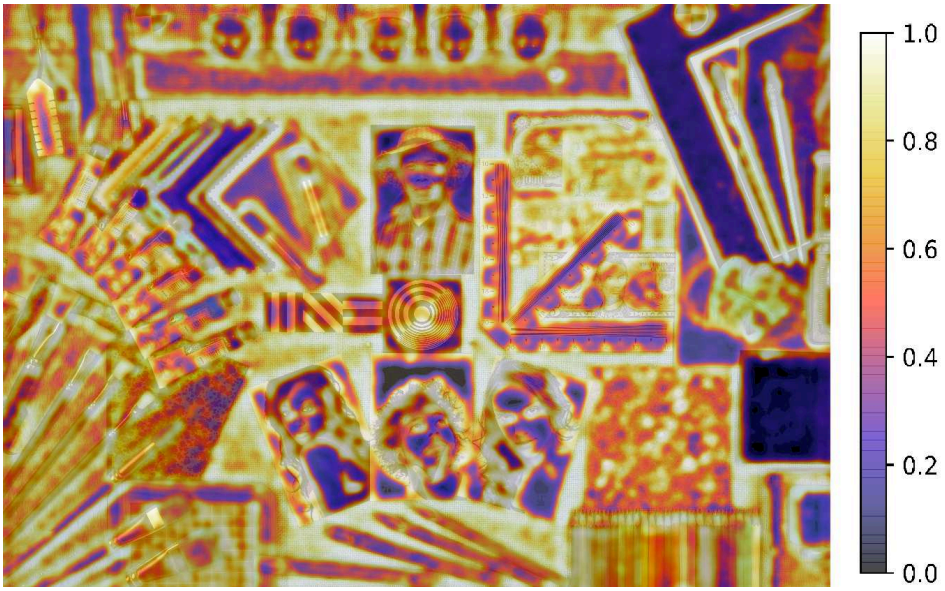


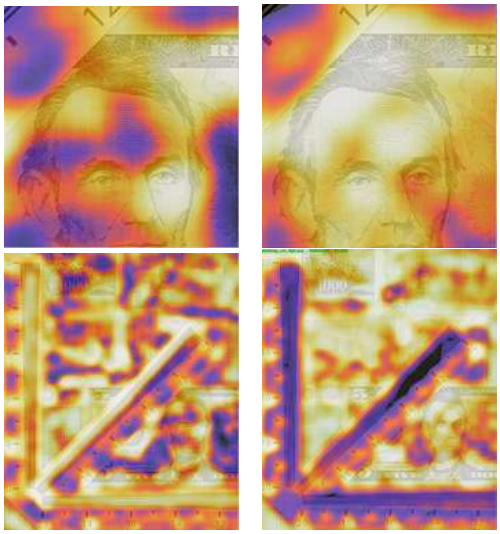
FIGURE 3.4: Normalized discriminant-region map S (better viewed with digital zoom). For display, we employ histogram equalization for normalization and obtain values from 0 to 1.

considered by our algorithm as the least discriminant for texture quality assessment. In particular, this is visible in the bottom-right regions on the black square patch. On the contrary,

regions with low contrast and many small details appear to be more discriminant (see around the banknote region). Results on wooden regions seem to depend on wood grain.

This analysis is performed considering all the images. We now propose to analyze the regions that discriminate devices among only low-quality or only high-quality images. For this analysis, the test set is split according to the ground-truth score. In this way, we compute two discriminant maps. Two small crops of these two maps are shown in Fig.3.5. Interestingly, we observe restricting our analysis to high-quality or lower-quality images leads to differences in results. For example, we observe that the resolution lines (in the bottom row of 3.5) discriminate for low-quality images, but not for higher-quality images. Conversely, areas exhibiting only very fine details are not the most useful for low-quality images. In particular, the forehead of the man is not discriminant among low-quality images, while this region is highly discriminant among high-quality images. It shows that the regions with very fine details are discriminant only among high-quality images since these details are completely distorted by all the low-quality devices.

Complete maps for several quality ranges are available in the appendix A



(A) Low quality images (B) High quality images

FIGURE 3.5: Comparison of discriminant-region maps for high-quality images and low quality. We display two patches extracted from the confidence maps obtained when using texture quality maps only from low (*ie.* Left) and high (*ie.* Right) quality images.

3.4.5 Comparison with state-of-the-art

Number of devices		20	60	100	140	20	60	100	140
Method	**Chart**	**SROCC**				**KROCC**			
RR Acutance [19]	*Gray-DL*	**0.704**	0.794	0.747	0.788	**0.533**	0.595	0.592	0.592
ResNet [53]	*Gray-DL*	0.641	0.795	0.792	**0.824**	0.464	0.598	0.592	0.630
DR^2S (Ours)	*Still-Life*	**0.830**	**0.912**	**0.890**	**0.900**	**0.638**	**0.740**	**0.716**	**0.728**

TABLE 3.2: Comparison of deep learning systems on different charts to [19]. *SROCC* and *KROCC* metrics are reported.

In this section, we compare the performance of our approach to existing methods. This comparison is twofold since both the methods and the charts need to be compared. We perform two sets of experiments. In the first set of experiments, we compare different methods on the two large datasets recorded with the *Gray-DL* and the *Still-Life* charts and the same 140 devices. The second set of experiments consists of a comparison of the devices on the *Color-DL* chart. This second set of experiments is highly challenging for learning-based methods because of the limited amount of training data.

Large database experiments

First, in our preliminary experiments, we observed that, for this experience, adding a small amount of Gaussian noise and random change in exposition leads to better performance. This data augmentation is performed on the fly on every training patch. Our main competitor is the *RR acutance* methods proposed in [19]. For the acutance computation, viewing conditions were set to 120 centimeters printing height and 100 centimeters viewing distance. Note that, the *RR acutance* method is intrinsically designed for the Dead-Leaves charts and cannot be used for the *Still-Life* chart. We include a second deep learning-based method for the *Gray-DL* chart in our comparison. This approach consists of a ResNet-50 [53] where the classification layer is replaced by 3 additional fully-connected layers and a linear regression layer. For this approach, we employ the *Random patch* strategy described in Sec. 3.4.3 inside of the texture region. Importantly, we do not report the performance of DR^2S on the

Gray-DL chart since the chart is designed to be uniformly discriminant for texture quality assessment.

Quantitative results are reported in Table. 3.2. First, we observe that with a limited number of devices for training (e.g. 20 devices), *RR Acutance* performs better than ResNet-50. However, the proposed approach clearly outperforms the texture-MTF based method ($+0.126$ and $+0.105$ in *SROCC* and *KROCC*, respectively). It shows that when few training samples are available, selecting the appropriate regions is essential for good performance. ResNet performance increases with the number of devices: with 140 devices, ResNet-50 clearly outperforms *RR Acutance* according to both metrics showing the potential of learning-based methods. While comparisons between results obtained using different charts must be interpreted with care, this result clearly shows that a learning-based approach can be intrinsically better than acutance-based methods using the exact same input images. Finally our DR^2S method on the *Still-life* chart leads to the best results according to both metrics and for every number of devices.

Small database experiments

Concerning the second set of experiments, we compared the different methods on the *Color-DL* chart. Note that the 14 devices of the *Color-DL* chart are a subset of the devices of the *Gray-DL* and *Still-Life* charts. Consequently, we also performed experiments using only these 14 devices on these two other charts. Note that this setting is very challenging for the two learning-based methods (ResNet and DR^2S) because of the limited amount of training data. Therefore, for the two learning methods, we perform two experiments. First, training and testing are performed using 14-fold cross-validation on the exact same data as the other methods. Second, we train our model on the complete database (without test devices) and test on the 14 devices in common with the *Color-DL* chart. These two variants are referred to as *Restricted* and *Full*. Again, we do not report the performance of DR^2S on the *Gray-DL* and *Color-DL* charts since the chart is designed to be uniformly discriminant. Results are reported in Table 3.3.

Method	Chart	SROCC	KROCC
FR Acutance [119]	*Color-DL*	0.701	0.544
RR Acutance [19]	*Gray-DL*	0.714	0.552
ResNet - Restricted	*Gray-DL*	0.640	0.463
ResNet - Full	*Gray-DL*	0.780	0.598
DR^2S - Restricted	*Still-Life*	0.746	0.569
DR^2S - Full	*Still-Life*	**0.873**	**0.702**

TABLE 3.3: State of the art comparison: Performance on the 14 devices database. Deep learning systems on perceptual and Gray-DL are compared to [19] and [119]

First, we observe that the MTF-based methods perform similarly on the color and gray-scale dead leave charts. It shows that the better performance of the proposed model on the *Still-Life* chart is not due to the lack of colors in *Gray-DL* but to its content. Second, using the restricted database, the ResNet method under-perform MTF-based predictions. However, when the amount of training data is sufficient, it outperforms *FR Acutance* and *RR Acutance*. Concerning DR^2S used with *Still-Life*, higher correlations are achieve even under data restrictions.

3.5 Extension to noise

We compare the measurements performed on the *DeadLeaves* chart and predictions on the *Still-Life* chart on the whole database (293 devices). We chose to benchmark our method to three different formulas of the visual noise metric:

- The formula standardized by CPIQ [1] (VN_{CPIQ})
- The formula in discussion in ISO15739 and lastly proposed in [137] (VN_{ISO})
- The formula used by DXOMARK [37] ($VN_{DXOMARK}$)

As the visual noise metric provides one metric for each patch, we consider for each formula the one interpolated for $CIE - L^* = 50$. Besides this, visual noise takes into account the

sensitivity of the human eye to different spatial frequencies under various viewing conditions. Hence the measurement is always dependent on the size of the image (i.e. print or on-screen) and the viewing distance. The effect of the viewing conditions is to stretch the CSF along the frequency axis. To evaluate the ability of the visual noise measure to assess the noise level in our dataset, we use two different conditions:

- Viewing Condition Print: a commonly used viewing condition of a print of 120 centimeters height viewed at 100 centimeters
- Viewing Condition Display: a viewing condition as the one used during the annotation process, involving a display viewed at 40 centimeters with a pixel pitch of 0.27 millimeters

Moreover, our method on the *Still-Life* chart, gives predictions on two areas of interest for each image: *Woman* and *Feather*. We will therefore evaluate the predictions of *Woman* and *Feather* compared to the ground truth of their respective areas as well as the average of the two predictions compared to the average of the annotations. Quantitative results are reported in Table 3.4

Method	**Viewing Condition**	**SROCC**	**KROCC**
VN_{CPIQ}	Print	-0.640	-0.460
VN_{CPIQ}	Display	-0.620	-0.445
VN_{ISO}	Print	-0.585	-0.416
VN_{ISO}	Display	-0.576	-0.408
$VN_{DXOMARK}$	Print	-0.646	-0.464
$VN_{DXOMARK}$	Display	-0.654	-0.470
Ours Zone 1		0.883	0.717
Ours Zone 2		0.862	0.689
Ours Average		**0.904**	**0.734**

TABLE 3.4: Performance on the devices database.

First, we observe that our method strongly matches the provided annotations and that it outperforms other benchmarked methods. These results must be carefully weighed since

the predictions were made on the same chart as the annotations (*ie.* the Still-Life chart), while the visual noise metrics were established on the Dead Leaves chart. We conclude that measuring the noise on uniformly gray patches does not sufficiently allow us to evaluate the perceived level of noise of the camera on a natural image.

3.6 Conclusion

In this chapter, we proposed a method that can estimate the perceptual quality of images. Our learning-based algorithm selects the chart region that is the most suitable for texture quality assessment. Our results also suggest that, if enough training samples are available, learning-based methods outperform MTF-based methods. Our method allows an accurate selection of a relevant region from a perceptual chart for each of these attributes. However, our method focuses on estimating image quality from a single chart in a laboratory environment. Consequently, in chapter 4 we focus on evaluating the texture quality for different scenes on natural images.

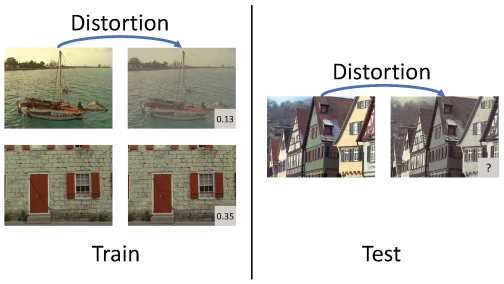
Chapter 4

Camera-quality assessment in real-world conditions

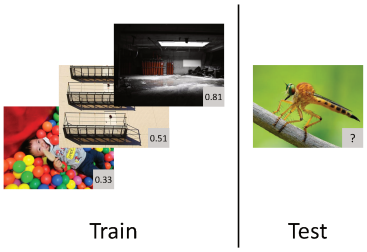
4.1 Introduction

When evaluating camera quality, several image attributes can be evaluated such as dynamic range, saturation, white balance, noise, or the presence of various artifacts [18]. This task has been traditionally addressed using synthetic charts recorded in labs [127, 19]. In particular, for texture preservation, the acutance metric, derived from the modulation transfer function, is generally used as described by Phillips *et al.* [97] and has been standardized by the IEEE organization [11]. This approach has been shown to be less correlated to human perception than learning-based methods in [127]. Contrary to these works and the previous chapter (Chapter 3), we propose to address this problem in real-world conditions on natural content. The goal of this work is to design a method that can evaluate and compare devices on the same natural content allowing content-specific ranking.

The camera quality assessment (CQA) problem is closely related to two well-known Image quality assessment (IQA) problems: *full-reference* (Figure 4.1a) and *no-reference* IQA (Figure 4.1b). In *full-reference* IQA, the goal is to evaluate an input target image with the help of a distortion-free version of this image. This task is mainly explored in the context of telecommunication applications where IQA is used to compare different compression and



(A) *Full-Reference* image quality assessment



(B) *No-Reference* image quality assessment

(C) *Reference-based* camera quality assessment

FIGURE 4.1: While previous approaches evaluate the quality of distorted images (Figure 4.1a) or authentic images with no reference available (Figure 4.1b), we propose to evaluate cameras using a high-quality reference (Figure 4.1c)

transmission pipelines. Consequently, works on *full-reference* IQA are limited to distortions that do not capture the complexity of modern camera systems that embed several non-linear post-processing steps. In contrast, we do not have access to distortion-free images in our case since camera quality is evaluated on natural content.

Reference-Based (Fig 4.1c) quality assessment. Camera quality assessment is also related to *no-reference* IQA which consists in estimating the image quality without knowing the noise-free image. However, we consider that we have multiple images of the same content allowing more accurate quality estimation. Note that differently from aesthetics quality assessment, which aims to provide a single quality score that captures many features not related to the camera device itself such as framing, the beauty of the object, or lighting conditions [120], we are not interested here in non-technical attributes of image quality.

With the rise of deep learning technologies, learning-based methods have been proposed

for IQA tasks [65, 14, 131, 153, 66, 142, 127]. The main advantage of learning-based approaches lies in their ability to predict quality scores that match human judgment. After collecting human preferences or ratings for a large set of images, a neural network can be trained to mimic human quality evaluation. Currently, existing public datasets for IQA can be divided into two categories: the images either present synthetic [114, 99, 98] such as blur or compression noise, or authentic [142, 143, 57, 38] distortions. On the one hand, synthetic datasets employ distorted images computed from pristine source images and a set of synthetic degradations. These synthetic datasets do not incorporate the types of complex distortions seen in authentically distorted images captured by mobile devices. Modern mobile cameras embed complex optical systems and algorithmic post-processing steps that result in images with complex distortions that are a mix of multiple distortions [41]. Furthermore, synthetic distortions classically used mostly refer to compression or transmission scenarios. Consequently, these datasets are not well-suited for camera quality assessment. On the other hand, authentic distortions datasets do present real-world distortions from actual devices, but every content appears only once or very few times in the dataset. This is problematic in the context of CQA since human annotators have difficulties rating image quality with changing content [96]. On the contrary, comparing devices on the same content allows content-specific ranking and more accurate annotation estimation [95].

To address these limitations, we introduce a new data collection protocol specific to the problem of camera quality assessment. We shoot the same natural scenes with many devices and estimate quality annotation using a procedure based on pairwise comparisons [95], combined with an efficient pair selection strategy [90]. The proposed protocol can be followed by any researcher or engineer who needs to learn their own CQA method. The resulting dataset provides a common test benchmark that can help the scientific community to develop new machine-learning algorithms with higher prediction accuracy.

In short, the contributions can be summarized as follows:

- We propose to address the problem of camera-quality assessment (CQA) in real-world

conditions on natural content. More precisely, we specify several variants of this problem and justify their respective practical advantages and limitations.

- We propose a data collection protocol tailored to evaluate learning-based camera-quality assessment methods. We collect 10 different scenes with many devices per scene with the same content. In every scene, a specific region of interest is selected. After a region alignment step, the images are shown to expert annotators and quality scores are estimated using a pairwise comparison protocol. Comparing devices on the same content allows us to obtain accurate quality scores [95]. We also take a photograph of every texture with a digital single-lens reflex (DSLR) camera that provides a high-quality texture reference for each scene. While a DSLR is not necessarily better than modern mobile devices concerning noise, color, or dynamic range, this is the case for texture detail preservation which is the evaluated attribute in this study. We propose to use this high-quality image both at annotation and evaluation times to improve the quality of the predicted scores. We refer to this setting as *reference-based* (Figure 4.1c) quality assessment. Our dataset is publically available [1].

- Finally, we benchmark several methods for the introduced CQA problems including baselines initially proposed for *no-reference* and *full-reference* IQA. This benchmark explores CQA for authentic distortions and assesses the performances of the algorithms both when the evaluated content is known or unknown. Our experiments allow us to extract several practical recommendations to ease the development of future CQA methods.

4.2 Related Work

While more detailed descriptions of datasets and methods can be found in Chapter 2, we propose in this section a brief review of datasets for image quality, as well as image quality

[1] https://corp.dxomark.com/carco-dataset/

assessment methods, whether *full-reference* or *no-reference* relevant to the motivation of this work or included in our experimental evaluation.

4.2.1 Existing datasets for image quality

Early datasets such as LIVE [114], CSIQ [73] or TIDs [98, 99]) are composed of noise-free images and subjective preference scores for several artificial distortions. These datasets mostly correspond to compression and transmission scenarios such as JPEG compression, sparse sampling, and reconstruction. The PieAPP dataset [101] also considers distortions from image processing algorithms, like deblurring, superresolution or reconstruction. Other datasets address the case of authentic distortions. For instance, the LIVE In the Wild [41] dataset consists of 1161 images shot with various devices with unique content. Similarly, the KonIQ10k [57] dataset contains images from a large public media database with unknown distortions. Yu et al. [143] collected a dataset of 12 853 natural photos from Flickr together with image quality annotations measuring the presence of several defects: exposition, white balance, color saturation, noise, haze, undesired blur. More recently, PaQ-2-PiQ [142] is a dataset comprising 40k real-world annotated images collected from social media and 120k annotated sub-patches. SPAQ [38] is a dataset dedicated to smartphone photography consisting of 11,125 pictures taken by 66 smartphones. While some of the contents are shared by several smartphones, this approach is not systematic and the information is not given on which images correspond to the which content.

None of the existing datasets present these two characteristics simultaneously: (1) the distortions are authentic and (2) the same contents are repeated many times. This justifies the creation of a novel dataset presenting these two characteristics, tailored for the need of camera quality assessment.

Most existing datasets are annotated through mean opinion scores. Annotators rate images and the image scores are averaged for all the annotations [57, 38]. However, Perez-Ortiz *et al.* [96] have shown that pairwise comparisons aggregated with a Thurstonian model

improve annotation quality over mean opinion scores for a given annotation effort time. Another advantage of this approach is that the distance between two annotation scores can be interpreted as a probability that a person will prefer one image over the other. The advantage of this pairwise annotation strategy has been proved in the case of synthetic distortions as in the PieAPP dataset [101]. Due to these advantages, we chose to employ the Thurstonian model to annotate our dataset. Furthermore, in contrast to alternative protocols based on crowd-sourcing [143, 142], our annotation process was conducted in a controlled environment with expert annotators to obtain high-quality annotations.

4.2.2 *Full-Reference* methods

Full-reference methods assume a pristine image used as a reference in order to evaluate a degraded image. Due to this requirement, this setting is used on datasets obtained by adding synthetic noise. This setting mostly corresponds to compression or transmission scenarios, or as a tool for image reconstruction or super-resolution. A simple Peak-Signal-to-Noise ratio could be considered a *full-reference* criterion but it poorly reflects human preferences. Other learning-free methods such as Structural Similarity Index (SSIM) [134], Multi-Scale SSIM (MS-SSIM) [136], and HaarPsi [104] are more complex and often correlate better with human preferences. The SSIM [134] is widely used albeit imperfect as a performance metric for reconstruction tasks and super-resolution [76]. It considers image degradations as perceived changes in its structural information. Concretely, the SSIM is a composite index comparing the luminance, contrast, and structure of the reference image and the degraded image through sliding windows. As an alternative, the Feature Similarity index (FSIM) [149] is an influential *full-reference* metric that combines two feature maps derived from the phase congruence measure and the local gradients of the reference with the degraded image to assess local similarities. HaarPsi metric [104] also compares images via an intermediate feature space but it uses simple Haar-Wavelet filters as a feature extractor. This approach is faster to compute than the FSIM.

In the last decade, we have witnessed the emergence of learning-based methods such as WaDIQaM-FR [14] or DualCNN [131]. Learning-based methods for *full-reference* IQA mostly rely on Siamese architectures. More precisely, a common backbone is used to process a reference image and the image under evaluation. Assuming a given backbone network, the main technical choice lies in the choice of the aggregation operation that is used to compare the feature maps resulting from the two images [130]. The aggregated feature map is then fed to a regression module which outputs an estimation of the perceptual quality of the input image under evaluation. For example, in the *full-reference* version of WaDIQaM [14], features of a Siamese network are aggregated with the difference and concatenation of features from the reference and degraded image. DualCNN [131] uses features from min-, average-, and max-pooling, of the two images, as well as differences between these two sets of features. Alternatively, LPIPS [150] normalizes the activations in the channel dimension, scales each channel by a learned weight, and then computes the Euclidean distance between the reference and degraded image. This process is done at several scales across the network and the distances are then averaged.

In general, these methods are not used for camera quality assessment as we lack a pristine image in existing datasets. However, we propose in this chapter a dataset with registered images and a high-quality sample for each content. We test these *full-reference* methods on our dataset even though a high-quality image is not a perfectly accurate reference. We also extensively evaluate different aggregation strategies for Siamese networks in the case of camera quality assessment.

4.2.3 *No-Reference* methods

Earlier methods for *no-reference* IQA are often referred to under the umbrella term Natural Scene Statistics (NSS), which assume that perceptual distortions can be measured as deviations of certain statistical properties. Until 2012 such features were extracted in the Wavelet domain or Discrete-Cosine Transform domain. Mittal et al. [92] then proposed

the BRISQUE model using features from the spatial domain, with better results. More precisely, these features are the horizontal, vertical, diagonal, and anti-diagonal products of the mean subtracted and contrast normalized pixels. Assuming Gaussian and anti-Gaussian distributions, the parameters of these distributions are then fed into a support vector regression. NIQE [93] uses the same handcrafted feature as BRISQUE, but instead of training a support vector regression compares the distribution of the patches of the test image to the distribution from a collection of natural pristine images. IL-NIQE [148] enriches the NIQE model with features related to colors, gradients, and the log-Gabor responses of the image. NIQE [93] and IL-NIQE [148] methods also are classified as NSS methods. While these methods predate the deep learning revolution, they are still used in benchmarks [38]. Moreover, we include these methods since their performance gap with deep-learning-based approaches is indicative of the non-triviality of the task and the dataset.

In 2014, Kang et al. [65] uses for the first time a convolutional network (CNN) for image quality assessment instead of handcrafted features. The proposed architecture is rather small: it takes as input 32x32 patches and the network depth is one single layer. Later works propose more complex architecture as WaDIQaM-NR [14]. WaDIQaM [14] proposes a convolutional network equipped with a patch weighting estimator. PaQ-2-PiQ [142] uses in addition to a convolutional network a Region-of-Interest pooling layer similar to the one employed in Fast-RCNN [42]. The question of the region of interest selection and weighting is indeed prevalent in the image quality literature [124, 25, 127]. Some works adopt probabilistic formulations modeling the distribution of the quality scores of each image. For example, a ConvNet is trained to predict opinion score histograms in [120] or Gaussian modeling is employed in [144] to capture the uncertainty of the annotated scores. Zhang *et al.*, with their DBCNN architecture [153], propose another *no-reference* method that is composed of two different convolutional models: the first one trained on synthetic distortion classification and the second is a pre-trained VGG [116] on ImageNet. This pre-trained model is introduced to better represent the perceptual features of natural images. The two sets of features are merged into a single representation for a final quality prediction with a bilinear pooling. Finally, while previous approaches were based on convolutional networks, MUSIQ [66] employs

the recent vision-transformer architecture. Their architecture ingests patches interpolated at different scales. The location token is computed from a look-up table common to every scale. The high architectural diversity of these methods motivates our thorough experimental evaluation where all these methods originally designed for IQA are compared in the context of CQA.

4.3 Dataset Collection

In this section, we describe our procedure to collect the proposed dataset for CQA. We refer to our dataset as ***Camera quality Assessment in Real-world COnditions*** dataset (or *CARCO*):

Image collection and pre-processing. The first step of the data-collection process consists in shooting different scenes, each taken by different cameras with different quality levels. In total, we shoot $D = 10$ different textures. The ten textures are selected from different scenes, except for one scene where we selected two regions since this particular scene contains diverse low-level details. On average, each scene is shot with 68 different devices. We employ camera devices with various qualities that are commonly embedded in standard smartphones, from various brands including but not limited to Apple, Google, Huawei, Xiaomi, Asus, OnePlus, Oppo, Sony, Vivo. To mimic real scenarios where devices can be added sequentially to the benchmark, we shoot the different textures at different dates and times of the day. Capturing images for multiple devices on a single day would indeed heavily bias the database toward very particular meteorological conditions. Consequently, the accuracy of the estimated scores would suffer from this data bias if a new test device is evaluated with different weather conditions. Capturing images on different days with varying conditions is a simple solution to mitigate this problem. Every picture from mobile devices is shot with default automatic settings. For each scene, we capture the reference image using a DSLR camera that produces images of superior quality. These images will be given to the annotators so that they can judge the authenticity of the details presented by the other images of each scene. We used a Panasonic DCS1R for all of our scenes. While DSLRs cannot

guarantee superior quality in terms of attributes such as noise or dynamic range, the large resolution of the sensor ensures better preservation of detail, which is our main focus in this study.

We choose scenes of various natures to capture the diversity of textures that may appear in pictures taken by camera users. Natural images usually contain various textures and we need to specify the image region of interest. Blurry backgrounds and uniform regions are not relevant for camera evaluation since every device would perform similarly. Thus, we select image regions with fine details so that the region is discriminative for texture rendering evaluation. The selected region sizes vary from 400×400 to 600×600 pixels on the DSLR image, while the original resolution of the DSLR camera is 8386×5584. Note that, this manual selection of the region of interest corresponds to a common evaluation procedure: the human annotator selects a region of interest to obtain a device evaluation that is not global but specific to the selected texture.

Collecting reliable quality scores calls for certain requirements. First of all, the content in the images shown to the annotators must be the same. To this aim, align every image with a reference crop. Note that, we only align the selected region of interest, and not the full images, since a global alignment of the entire content would lead to inferior registration accuracy on the regions of interest. To guarantee accurate alignment, we successively test several alignment approaches and visually inspect the result. This procedure is more cumbersome but it allows much greater alignment than a fully automatic pipeline. To achieve a satisfying trade-off between long manual labor and alignment quality, the choice of these alternatives is done once for all the images of each scene. More precisely, we proceed as follows.

- For some scenes, we estimate depth maps from the input images and identify the image regions with an object-to-camera distance that matches the distance estimated in the reference patch taken with the DSLR. We use publicly available pre-trained convolutional depth map monocular estimation networks [7]. We then select with manual thresholds on standardized depth maps to select for our reference shot containing our

region of interest. We observed that this initial cropping stage is useful in cases where the input scene is multi-planar.

- We extract AKAZE [6] descriptors helped by a scene-specific model constituted by manually selecting several remarkable points in a few images and then estimate the optimal homography using the SCRAMSAC [110] model selection algorithm.
- For some scenes, the AKAZE descriptors do not perform well. In these cases, we employ the template matching technique described in [24].

A quantitative evaluation of our registration pipeline is highly difficult since ground-truth registration parameters are not available for real-world images. To illustrate the quality of our registration pipeline, in Figure 4.2, we display ten random patches of the same content. Overall, this visual inspection confirms that the patches are well-aligned.

FIGURE 4.2: Images samples from one of the scenes. It illustrates the accuracy of the image registration pipeline.

As a second requirement, the content must be presented at the same scale for the annotator to judge with the same visibility. Images are rendered at different resolutions ranging from 10 to 46 megapixels for the DSLR. Most mobile devices render images at about 12 megapixels. We interpolate the content to the largest image size for each scene, as downscaling any image would result in a loss of information. In this step, we preferred the bicubic interpolation to nearest-neighbor and bilinear interpolations since it produces fewer artifacts.

Annotation.

Once the images are aligned and scaled, we proceed to their annotation. Our annotation process is based on pairwise comparisons. Indeed, pairwise comparisons are easier for a human annotator than estimating absolute values by observing an image alone [96]. To obtain a satisfying trade-off between annotation time and reliability of the resulting scores, we chose to compare few image pairs and estimate quality scores for every image from these sparse comparisons. The annotators were instructed to select the image with the highest level of detail, regardless of other impairments. The process for the 10 scenes takes about one hour per annotator. We employ 22 annotators and combine a total of 29297 pairwise comparisons using the algorithm proposed by Perez-Ortiz *et al.*[96]. This algorithm employs a Thurstonian model [12] to estimate continuous scores with scale homogeneity in terms of perceptible difference. Using a Bradley-Terry model is an alternative, where a logistic function is used instead of a Gaussian in order to be more computationally effective. We did not investigate this alternative as this leads to similar solutions [95]. In this model, the scores are modeled as random variables with $\boldsymbol{r}$ the vector of samples scores described as $r_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$ with $i \in [1...N]$, where N denotes the number of samples. Thus, the probability that an annotator choose an image i better as the image j is $\Phi\left(\frac{\mu_i - \mu_j}{\sqrt{2}\sigma_{ij}}\right)$ with Φ the cumulative distribution function of the normal function, with $\sigma_{ij}^2 = \sigma_i^2 + \sigma_j^2 + \beta^2$, with $\beta = 1$ describing the annotator noise. The scores $\boldsymbol{r}$ can be computed from the set of comparisons using this statistical model with a maximum likelihood estimation as detailed in [111]. To robustly estimate quality scores from a pairwise annotation experiment, image pairs to be compared have to be chosen wisely. Indeed, proposing trivial pairs to an annotator brings little information on the respective quality of each image. On the contrary, annotating pairs with similar qualities leads to more accurate scores with fewer comparisons. In order to ask the annotator his preference for pairs that are the most useful to the annotation process, we compute the information gain for each sample pair following [90]. We compute the Kullback-Leibler divergence between the current and posterior distributions of $\mathbf{r}$ for every pair possible of image samples, summed for each of the two outcomes of the comparison weighted by the event probability. The pair with the highest value is then selected. For better computational efficiency, the approach of [78] is employed to produce the next comparisons

as batches instead of computing the information gains at each comparison.

Analysis of the dataset.

Following this protocol, we obtain the dataset illustrated in Fig. 4.3. For every scene, we show the reference images, a medium-quality device, and a low-quality device. We can see that this dataset is fundamentally different from existing datasets: on contrary to synthetically generated datasets, our dataset presents authentic distortions from existing camera devices. The distribution of each scene is available in the fourth row of the figure. While a distance of 1 unit corresponds to a 75 % likelihood of preference, in most cases, scores span between 8 to 10 units. Our dataset differs from other in-the-wild datasets which either do not present the same content at varying quality [57, 41] or lack the step of region selection and registration [38] which is necessary to benchmark several camera devices. Furthermore, they do not provide reference images for each scene.

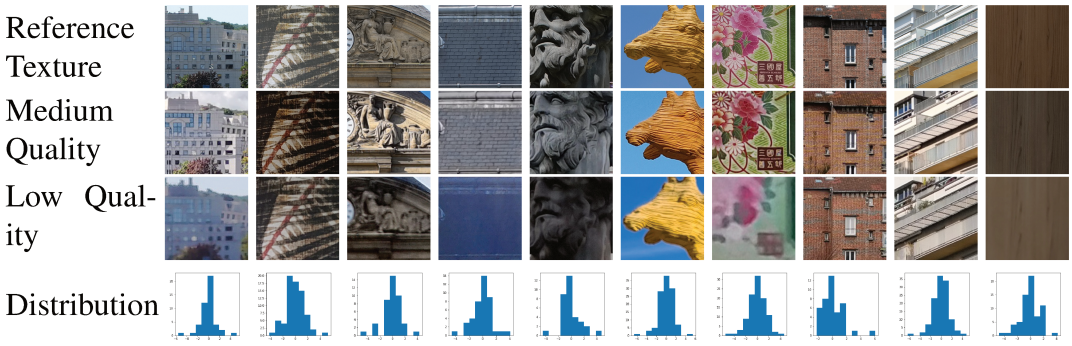


FIGURE 4.3: Images samples from the CARCO dataset. We present the different visual contents at varying quality with the corresponding score histograms. The images are better viewed with digital zoom.

In this analysis, we propose to evaluate score uncertainties. The standard deviations used in our Thurstonian model could be used to evaluate uncertainty but this solution may be biased since it uses our model assumption to estimate the uncertainty scores. Therefore, we prefer using an external procedure that does not rely on our model assumption. We use a bootstrapping procedure [59, 89] that is commonly adopted when assumptions of a parametric model are in doubt. This statistical procedure employs a re-sampling strategy to obtain multiple sets that follow the distribution of our observations. More precisely, our set of

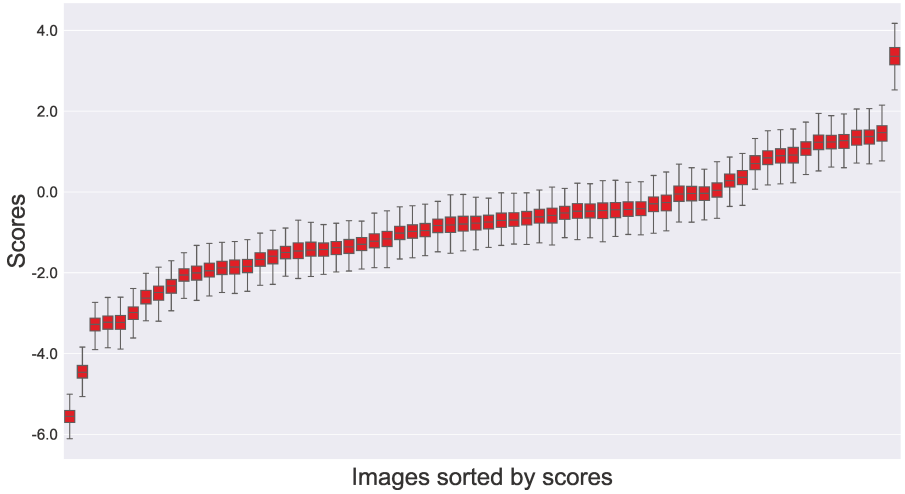


FIGURE 4.4: Score uncertainties estimated via bootstrapping. The population of scores for an image is then represented using a box-plot indicating the median, the first and the third quartiles, and the full range of the distribution.

pairwise comparisons is re-sampled 500 times with replacement: comparisons are sampled from the original set of comparisons until the population in the simulated set reaches the original set of comparisons size. The obtained sets are then converted to scores using our Thurstonian model to obtain confidence intervals for each score. The result of this protocol is shown in Figure 4.4 for one of the scenes. In this Figure, we adopt the widely-used Tukey's range test to remove the outliers [125] to obtain a more robust visualization of the score uncertainties. Overall, we can observe that all the scores have relatively small ranges of uncertainty. We can also note that the ranking of the devices with extreme scores (very low or high quality) is much more certain than with medium-range scores.

4.4 Evaluation protocols

We now detail the evaluation protocol that we employ to benchmark existing methods on our CQA dataset.

4.4.1 Scenarios and settings

In our experiments, we adopt two different scenarios where the content of the test image is either known or unknown.

Known content. In this first Scenario, we allow semantic overlap between train and test sets. This scenario can be considered easier since test images depict content that has been seen at training time. However, the camera device evaluated at test time is still unknown.

This scenario corresponds to the use case where pictures of the scene were taken and annotated, but we wish to evaluate new devices on these contents. 70% of images from all scenes are designated as train scenes for each run. The metrics are then obtained on the remainder 30%.

Unknown content. In this second scenario, we impose no semantic overlap between the train and test set. This corresponds to the use case where we wish to evaluate camera quality on new scenes unseen during training and where no annotations have been done for this new content. Thus, the training set corresponds to all scenes but one used as a test set. This scenario is particularly useful to the generalization ability of the different models. Note that, the *unknown content* scenario is more classic in the IQA literature. Existing datasets either present images with unique content or consist of distorted images which render the *known content* scenario irrelevant and trivial. However, since our goal is to evaluate the camera itself, it is practically relevant to assume that the experimenter can evaluate the test device on the same content as the training images.

Settings. In both scenarios, we conduct experiments in two settings: with the help of a high-quality image and without. These two settings are referred to as *reference-based* and *no-reference* CQA, respectively. *Reference-based CQA* is closely related to *full-reference* image quality assessment. However, in our case, the reference image is not a distortion-free image but an example of an image taken by a high-quality device.

In short, we have two scenarios and two settings, leading to four possible combinations. We now describe the different requirements and practical use cases for an end-user in every possible case:

- *Unknown content* and *no-reference*: this is the least constraining case for a user. Our dataset can be used directly: the user can simply take a single picture of any new content with the test device. The image can then be given to a model trained on our dataset to obtain the estimated score. Note that, we also release pre-trained models to facilitate experiments.

- *Unknown content* and *reference-based*: compared to the previous case, a high-quality image of the texture must be added as an input to a *Reference-based* model. This setup also requires little effort, as a user can still use our pre-trained models.

- *Known content* and *no-reference*: in this case, the user has to repeat the data collection and annotation process using multiple devices on the content of interest. Then, this newly collected content can be added to our dataset in order to train a model. The model can then be used in the future to evaluate any new device. It requires more effort for this use-case compared to previous ones, but having an understanding of the content help a more accurate quality estimation. Nevertheless, we argue that this setting is less practically relevant than the others since, the effort of including a DSLR among the set of devices is negligible compared to the total data-collection process. In this scenario, the *reference-based* case is much more practically relevant and consequently, we retain this combination in our experimental evaluation only for the sake of comparison.

- *Known content* and *reference-based*: Similarly to the previous case, the user also needs to collect images of the new content with multiple devices. The only difference is that the content must also be shot with a DSLR camera to obtain a reference image. Therefore, this use-case is not significantly more constraining than the previous one but offers a more accurate evaluation of quality, as seen in Section 4.5.

4.4.2 Metrics

Mean absolute error is not classically used in the image quality literature [66, 142]. It does not allow the comparison of methods whose outputs are in different value ranges. Furthermore, as in our case, the scenes are not cross-content calibrated, we are more interested in correlations between ground truths and prediction. Following previous works [92, 93], we use the linear correlation coefficient (LCC), also known as Pearson's linear correlation coefficient between annotations and predictions. Additionally, we rely on two distinct metrics based on the correlation of the rank order. First, we adopt the Spearman Rank-Order Correlation Coefficient (*SROCC*) defined as the linear correlation coefficient of the ranks of predictions and annotations. Second, we report the Kendall Rank-Order Correlation Coefficient (*KROCC*) defined by the difference between concordant and discordant pairs divided by the number of possible pairs. The three metrics are estimated independently for each scene since the annotations are not calibrated across scenes. Every experiment is run five times and we report the results averaged over the 10 scenes and these five runs. To evaluate the impact of stochasticity at training time, we also report the standard deviation over the five runs of the average score.

4.4.3 Evaluated baselines

In addition to the existing IQA approaches described in section 4.2, we propose to add several other experiments for the referenced-based setting. A siamese architecture uses a common backbone to process the reference image and the image to be evaluated. The backbone outputs a feature representation for each image, which is aggregated to perform the task with information from both images. Since the main technical point referring to the use of siamese networks is the aggregation operation between features of the two images, we propose in this chapter to look at the performances of several aggregation schemes. The features obtained after the aggregation are then fed in a two-layer regression head. Figure 4.5 represents the siamese architecture and the role of the aggregation layer. Both the reference patch and the evaluated images are fed into a convolutional network utilized as a feature extractor.

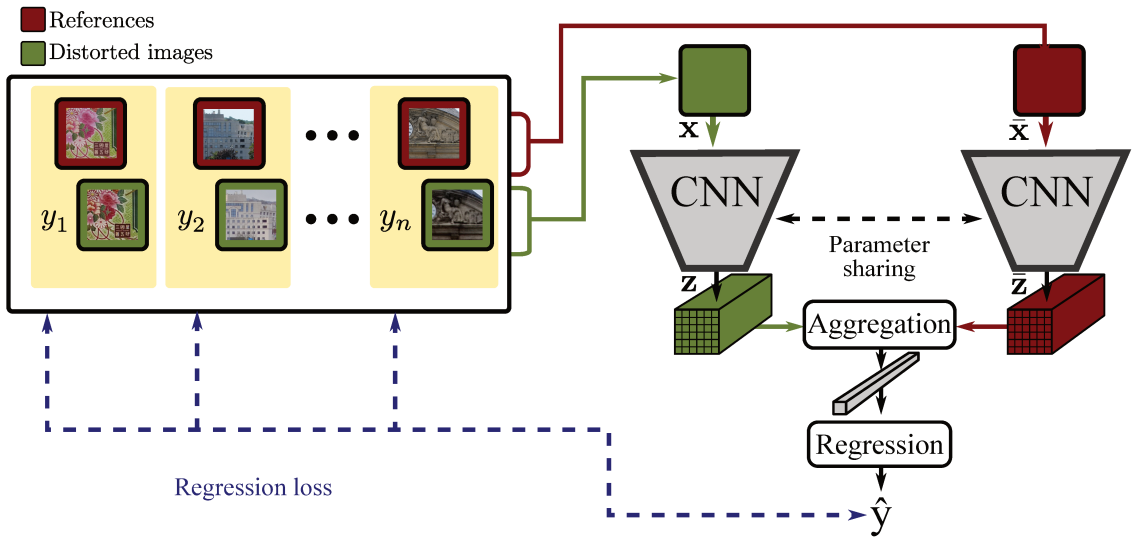


FIGURE 4.5: Pipeline overview for the introduced *reference-based* baselines. Images from different scenes are paired with the corresponding reference and given as inputs to a siamese network. Feature maps are aggregated into a 1-D vector and supplied to a regression head.

The network outputs feature maps which are combined into a feature vector via the use of an aggregation layer. A regression head predicts a score for the input image, which is then compared to the ground-truth scores with a regression loss function at training time to update the network's weights.

For the *reference-based* experiments, features estimated by both branches of the Siamese network are aggregated in different layers from the network in order to compare the image at different scales. For the VGG experiments, the selected layers are the same as in the LPIPS [150] method. For the ResNet-18 experiments, the selected features are located after each ResNet block and correspond to different scales (112x112, 54x54, 28x28, 14x14, 7x7). We call this approach "feature pyramid" in accordance with the literature [81]. The extracted features at each resolution are then concatenated and given to a linear regression layer.

We now describe the layers that we introduce to compare the reference image with the image to evaluate. We consider an image $\mathbf{x}$ with its corresponding reference $\bar{\mathbf{x}}$. Building upon the siamese architecture [69], we consider a backbone convolution network ψ and map both images with feature maps $\mathbf{z} = \psi(\mathbf{x}) \in \mathbb{R}^{H' \times W' \times C}$ and $\bar{\mathbf{z}} = \psi(\bar{\mathbf{x}}) \in \mathbb{R}^{H' \times W' \times C}$.

We aim at comparing $\mathbf{z}$ with $\bar{\mathbf{z}}$ in order to evaluate the quality of the image corresponding

to $\mathbf{z}$. To compare the two tensors, we treat separately each channel and consider that the two feature maps provide $H' \times W'$ samples of two random variables.

We note respectively $\mathbf{z}$ and $\bar{\mathbf{z}}$ z_{ijc} and $\bar{z}_{ijc}$, $\forall (i,j,c) \in H' \times W' \times C$. We will also note $\mathbf{z}_c$ and $\bar{\mathbf{z}}_c$ these tensor at channel indice c, and z_c and $\bar{z}_c$ their mean.

$\mathbf{L_1}$. Classically, the L_1 distance can be used as aggregation. In this case, the output vector $\boldsymbol{\alpha} = \big(\alpha_c\big)_{1 \leq c \leq C}$ is expressed as

$$\alpha_c = \mathrm{GAP}(|\mathbf{z}_c - \bar{\mathbf{z}}_c|) \tag{4.1}$$

Concatenation. The concatenation aggregation is defined as

$$\alpha_c = \mathrm{GAP}(\mathbf{z}_c) \cdot \mathrm{GAP}(\bar{\mathbf{z}}_c) \tag{4.2}$$

where GAP is the global average pooling layer. Note that in this case, the dimension of $\boldsymbol{\alpha}$ is $2C$. Concatenating is a very natural way to combine information and is widely used in the literature with siamese networks

Both the concatenation aggregation and the L_1 aggregation do not exploit the spatial structure of the feature map. While the L_1 explicitly computes a distance between the image to evaluate and the high-quality image, the concatenation relies on the regression head to process and aggregate the information.

Convolution. We provide an additional baseline where feature maps are concatenated and processed by an additional convolutional layer, which seems natural for a convolutional network, equipped with an activation layer and batch normalization. The resulting features can be expressed as

$$\alpha_c = \mathrm{GAP}(\mathrm{Conv}(\mathbf{z} \cdot \bar{\mathbf{z}})) \tag{4.3}$$

Image quality metrics as aggregation. Following [130], we propose to use full-reference

quality metrics in order to aggregate feature maps from the reference and the image to evaluate. Using previously introduced notations, we have $\boldsymbol{\alpha} = \left(\alpha_c\right)_{1 \leq c \leq C}$ with

$$\alpha_c = \mathrm{IQM}(\mathbf{z}_c, \bar{\mathbf{z}}_c) \tag{4.4}$$

with IQM an image quality metric, in this chapter, we tested SSIM [134] and HaarPsi [104] metrics. These methods have the advantage of exploiting the spatial relationship between different features of the feature map

Correlation. We propose to measure the linear dependence between the two random variables by computing their correlation. The output vector $\boldsymbol{\alpha}$ of our correlation aggregation layer is given by:

$$\alpha_c = \frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \frac{(\mathbf{z}_c - z_c)\,(\bar{\mathbf{z}}_c - \bar{z}_c)}{\sigma_c \bar{\sigma}_c + \epsilon} \tag{4.5}$$

where ϵ is set to 10^{-5} for numerical stability and z_c and $\bar{z}_c$ denote the mean value for the channel c:

$$z_c = \frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} z_{ijc} \text{ and } \bar{z}_c = \frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \bar{z}_{ijc}.$$

where σ_c and $\bar{\sigma}_c$ denote the standard deviation for the channel c:

$$\sigma_c = \sqrt{\frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} (z_{ijc} - z_c)^2} \tag{4.6}$$

and

$$\bar{\sigma}_c = \sqrt{\frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} (z_{ijc} - \bar{z}_c)^2}. \tag{4.7}$$

This method does not exploit the spatial aspect of the feature map, though it computes a second-order moment between the two variables.

4.4.4 Implementation details

We implemented our framework using the Pytorch library. We employ the Adam optimizer [67] using the default parameters. For every experiment, the models are trained for 60 iterations. We use the Huber loss with a 0.3 threshold. For models using a standard vision backbone, the regression head is composed of two fully connected layers, with the hidden layer having 256 neurons.

4.5 Experiments

We now present the results of our experimental benchmark starting with *reference-based* CQA methods. Then, we evaluate the different methods for *no-reference* CQA.

4.5.1 *Reference-based* CQA

This set of experiments is organized in three sections. First, we evaluate non-learning-based approaches that address both scenarios (*known* and *unknown content*) identically. Then, we compare learning-based approaches in the *known* and *unknown content* scenarios.

Non-learning based Methods

First, we evaluate off-the-shelf methods, that either do not require any training *SSIM* [134], *MS-SSIM* [136], *FSIM* [149], and *HaarPsi* or are pre-retrained on an external dataset (*LPIPS* [150]). To compare these methods with learning-based methods, we consider a simple baseline based on ResNet-18 equipped with a correlation aggregation layer. We provide two variants of this baseline: in the *known content* scenario, images of the test scenes are included in the training set (but the device are different between training and test), while in the *unknown content* scenario, we assume no scene-overlap between training and test. As detailed in Table 4.1, non-learning-based methods perform rather poorly: the best performance is

Method	LCC	SROCC	KROCC
SSIM [134]	0.484	0.315	0.228
MS-SSIM [136]	0.396	0.231	0.168
FSIM [149]	0.533	0.358	0.264
HaarPsi [104]	0.644	0.474	0.345
*LPIPS** [150]	**0.807**	**0.763**	**0.608**
Correlation Agg. Unknown content †	0.918	0.898	0.755
Correlation Agg. Known content †	**0.943**	**0.917**	**0.803**

TABLE 4.1: Experiment in the *Reference-based* setting with off-the-shelf methods. *uses a pretrained model. †uses a ResNet-18 backbone.

achieved by the HaarPsi [104] method, which only achieves 0.644 for the linear correlation, followed by the FSIM [149]. Interestingly, the MS-SSIM [136] performances are worse than the vanilla SSIM [134] on our CQA dataset. However, LPIPS [150], even with no retraining on our CQA dataset, vastly outperforms non-learning-based methods. We can conclude from Table 4.1 that learning-based methods clearly outperform classical IQA metrics. These IQA metrics perform worse than a vanilla learning-based method with a ResNet-18 architecture, whether on *known* or *unknown content*. Interestingly, we can also notice that the three adopted metrics consistently lead to the same method ranking.

Known content Scenario

In Table 4.2, we compare the different methods in the *reference-based* setting in the *known content* scenario. In all these experiments we adopt a ResNet-18 backbone. This choice is later evaluated in a dedicated ablation study. To measure the gain brought by the availability of the High-quality image, we also include a vanilla *no-reference* ResNet-18 that takes a single image as input.

First, we observe that all the methods outperform the non-learning-based approaches compared in Table 4.1. This confirms the superiority of learning-based approaches for CQA. We also observe that in the *known content* scenario, it is possible to achieve very high performance compared to off-the-shelf methods.

Method		LCC	SROCC	KROCC
WaDIQaM-FR [14]		0.869 ± 0.019	0.846 ± 0.023	0.698 ± 0.025
LPIPS [150]		0.863 ± 0.033	0.806 ± 0.029	0.673 ± 0.025
DualCNN [131]		0.935 ± 0.007	0.927 ± 0.012	0.817 ± 0.020
Backbone	**Aggregation**	**LCC**	**SROCC**	**KROCC**
ResNet-18	L_1	0.894 ± 0.049	0.876 ± 0.024	0.762 ± 0.020
ResNet-18	Convolution	0.789 ± 0.040	0.748 ± 0.020	0.606 ± 0.024
ResNet-18	Concatenation	0.936 ± 0.008	0.912 ± 0.016	0.801 ± 0.021
ResNet-18	Correlation	0.943 ± 0.008	**0.917** ± 0.009	**0.803** ± 0.010
ResNet-18	SSIM	**0.945** ± 0.006	0.913 ± 0.021	0.796 ± 0.025
ResNet-18	HaarPsi	0.939 ± 0.009	0.915 ± 0.018	0.804 ± 0.020
ResNet-18	LPIPS-Like	0.924 ± 0.024	0.892 ± 0.016	0.762 ± 0.025
ResNet-18	*No-reference*	0.925 ± 0.006	0.899 ± 0.023	0.783 ± 0.028

TABLE 4.2: *Reference-based* experiments on *known content*.

We can also notice that the *no-reference* baseline outperforms several methods that were trained in the *known content* scenario. For instance ResNet-18 in *no-reference* setting outperforms L1, convolution, and LPIPS in terms of LCC. This shows that the choice of the aggregation layer is crucial to benefit from the availability of the reference.

These experiments also show that it is possible to achieve satisfactory performances without a high-quality image as a reference, even though improvements are seen in the case of the SSIM (+0.014 in SROCC), HaarPsi (+0.016 in SROCC), and Correlation aggregation (+0.018 in SROCC). The SSIM aggregation provides the best performance in terms of LCC while the best SROCC and KROCC is obtained by the Correlation aggregation.

Unknown content Scenario

Main Results. We now evaluate the different methods in the *unknown content* scenario. We evaluate four learning-based *full-reference* methods from the literature: DualCNN, a recent method, WaDIQaM, a classic image quality metric, and LPIPS, which is commonly used as an evaluation metric or as reconstruction loss. We also include PieAPP, a method

Method		LCC	SROCC	KROCC
WaDIQaM-FR [14]		0.854 ± 0.025	0.799 ± 0.035	0.634 ± 0.044
PieAPP [101]		0.728	0.645	0.490
LPIPS [150]		0.908 ± 0.009	0.882 ± 0.015	0.729 ± 0.020
DualCNN [131]		0.891 ± 0.009	0.875 ± 0.006	0.720 ± 0.007
ResNet-18	L_1	0.855 ± 0.011	0.816 ± 0.011	0.650 ± 0.011
ResNet-18	Convolution	0.745 ± 0.014	0.654 ± 0.012	0.492 ± 0.012
ResNet-18	Concatenation	0.904 ± 0.006	0.881 ± 0.006	0.730 ± 0.008
ResNet-18	Correlation	$\mathbf{0.918} \pm 0.004$	$\mathbf{0.898} \pm 0.003$	$\mathbf{0.755} \pm 0.005$
ResNet-18	SSIM	0.912 ± 0.005	0.893 ± 0.005	0.745 ± 0.007
ResNet-18	HaarPsi	0.906 ± 0.006	0.878 ± 0.09	0.724 ± 0.10
ResNet-18	LPIPS-like	0.887 ± 0.006	0.865 ± 0.005	0.705 ± 0.006
ResNet18	*No-reference*	0.870 ± 0.007	0.851 ± 0.007	0.694 ± 0.008

TABLE 4.3: *Reference-based* experiments on *unknown content*.

trained using pairwise preferences. However, the source code for their specific training procedure was not available, and therefore, we report the performance of the model trained on their IQA dataset. We also evaluate different variants of the siamese architecture equipped with the different aggregation described in 4.4.3. Results are reported in Table 4.3. In this more challenging scenario, the siamese architecture that receives the high-quality reference as second input improves significantly the performance on our dataset with respect to the *no-reference* baseline (+0.048 in LCC for the correlation aggregation). Conversely, with the *known content* setting, correlation aggregation obtains the best performance according to the three metrics while SSIM also consistently outperforms the HaarPsi aggregation method that was performing well in the *known content* scenario [2]. As in the *known content* scenario, the L_1 and convolution aggregations do not perform well. The *reference-based* version of WaDIQaM and DualCNN obtain results better than non-learning approaches (see Table 4.1) but under-performs most ResNet-18 based methods. The DualCNN method obtains performances close to the ResNet-18 baselines but still performs slightly worse. The PieAPP baseline, trained on synthetic distortions performs poorly. It shows that their synthetic distortions

[2] A pre-trained model of the ResNet-18 equipped with the correlation aggregation model trained on all scene of our dataset will be released upon acceptance

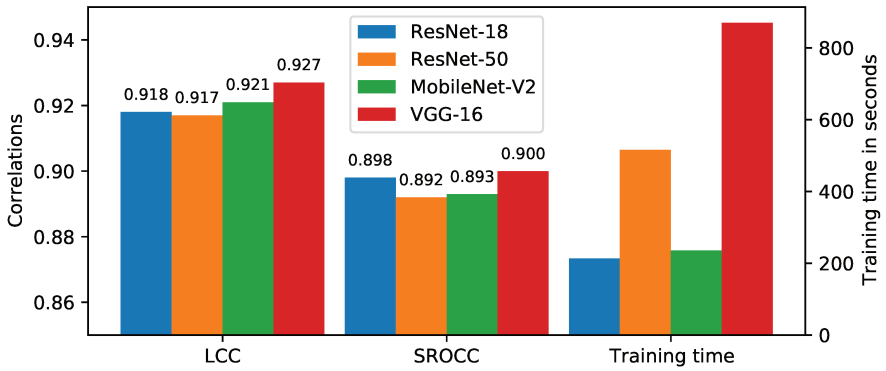


FIGURE 4.6: *Reference-based* backbone ablation experiments on *unknown content*. The correlation aggregation is used for this comparison.

differ from the authentic ones of our dataset.

In this *reference-based* setting, we now test different backbones to assess their influence on the quality of predictions. We choose two sizes of ResNet backbones (*ie.* 18 and 50) since ResNet architecture is widely used in computer vision literature. We also test the VGG-16 architecture, since it has been shown to extract relevant features for the human perception [76, 150]. Finally, we include the MobileNet-V2 architecture since it is a more computation-efficient alternative to ResNet and VGG. In all cases, we employ backbones pre-trained on ImageNet and adopt a correlation layer to aggregate the features coming from each input image since it provided the best results in 4.3. In these experiments, we employ the feature pyramid framework previously described in Sec. 4.4.3. Results are reported in Figure 4.6. The VGG-16 backbone is the best performer in all metrics. However, it should be noted that VGG-16 requires four times more training time than ResNet-18. The results show that the performance changes when changing the backbone network. However, we observe an amplitude of the results of only 0.10 in LCC, and 0.08 in SROCC. These differences are smaller than the differences due to the impact of the aggregation layer evaluated in Table 4.2.

Impact of the feature pyramid. Performing the aggregation at several stages of our backbone allows us to add features from different scales for the subsequent regression layer.

While this approach increases computation time and the number of parameters of the network, his approach is commonly used in IQA methods [66, 130]. We conduct an experiment to measure the importance of this multi-scale design. We adopt a siamese architecture with a ResNet-18 backbone. In Table 4.4, we measure the impact of the pyramid of features improvement with two different aggregations: Correlation and SSIM. We observe that the pyramid substantially improves the performance in the case of the correlation aggregation (+0.028 SROCC) and dramatically in the case of the SSIM (+0.113 SROCC) aggregation. Interestingly, while roughly equivalent in the case of a feature pyramid, the correlation aggregation vastly outperforms the SSIM aggregation in the no-pyramid case. Consequently, we strongly recommend the use of a pyramid architecture, especially since it increases training time by 20 % only.

Aggregation	Pyramid	**LCC**	**SROCC**	**KROCC**
Correlation	✗	0.901 ± 0.011	0.871 ± 0.009	0.717 ± 0.009
	✓	$\mathbf{0.918} \pm 0.004$	$\mathbf{0.898} \pm 0.003$	$\mathbf{0.755} \pm 0.005$
SSIM	✗	0.813 ± 0.009	0.780 ± 0.008	0.614 ± 0.013
	✓	0.912 ± 0.005	0.893 ± 0.005	0.745 ± 0.007

TABLE 4.4: Impact of the pyramid aggregation scheme on *unknown content* Scenario in the *Reference-based* setting. The ResNet-18 architecture is used.

Impact of the number of training scenes. We propose here a further analysis of the dataset and the performances of our method in a lower data regime. For this experiment, we choose the architecture with a ResNet backbone and a correlation aggregation scheme. For each scene taken as the target scene, we sample randomly a fixed number of training scenes. This process is repeated five times. As seen in Figure 4.7, steady improvements can be seen when training scenes are added for the result on the remaining scene. This is explained by additional data as well as additional diversity in the dataset.

Scene by scene detailed results We now report the detailed results of the four best-performing architectures providing SROCC scores for every scene. While HaarPsi performs best in the *known content* scenario, it underperforms SSIM in the *unknown content* scenario

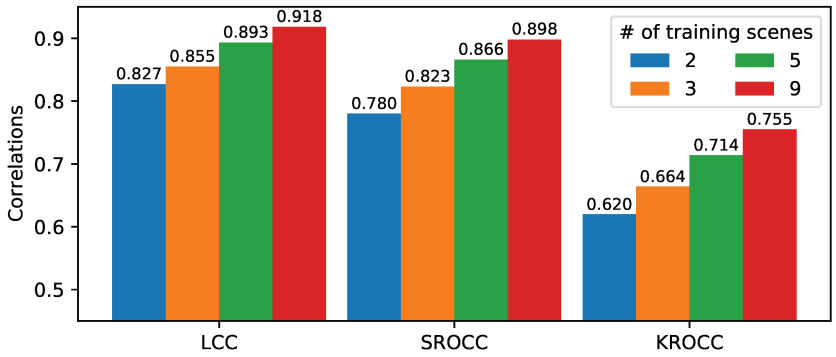


FIGURE 4.7: Impact of number of training scenes for *Reference-based* setting on *unknown content*. The correlation aggregation is equipped on a ResNet-18 for this comparison.

where a higher generalization capacity is required. The detailed results reported in Table 4.5 provide an explanation for this poorer performance of HaarPsi. While the HaarPsi performances are similar to the SSIM and Correlation aggregations for most scenes, HaarPsi fails in a few scenes such as #5 #6 and #10. We observe that every method witnesses a similar variation of performance from one scene to another.

Aggregation	**Scene #1**	**Scene #2**	**Scene #3**	**Scene #4**	**Scene #5**	**Scene #6**
Concatenation	**0.863** $\pm$ 0.019	0.867 $\pm$ 0.010	0.973 $\pm$ 0.005	0.939 $\pm$ 0.012	0.844 $\pm$ 0.015	0.836 $\pm$ 0.015
Correlation	0.842 $\pm$ 0.006	**0.893** $\pm$ 0.011	**0.975** $\pm$ 0.006	**0.949** $\pm$ 0.011	**0.874** $\pm$ 0.016	0.847 $\pm$ 0.012
SSIM	0.844 $\pm$ 0.017	0.882 $\pm$ 0.015	0.967 $\pm$ 0.004	0.896 $\pm$ 0.033	0.866 $\pm$ 0.019	**0.850** $\pm$ 0.012
HaarPsi	0.847 $\pm$ 0.023	0.868 $\pm$ 0.017	0.973 $\pm$ 0.004	0.905 $\pm$ 0.019	0.823 $\pm$ 0.026	0.812 $\pm$ 0.053
Aggregation	**Scene #7**	**Scene #8**	**Scene #9**	**Scene #10**	**Average**	**Inter-scenes std**
Concatenation	0.946 $\pm$ 0.005	0.847 $\pm$ 0.022	0.775 $\pm$ 0.025	0.939 $\pm$ 0.017	0.883 $\pm$ 0.006	0.055
Correlation	0.973 $\pm$ 0.001	0.857 $\pm$ 0.004	**0.834** $\pm$ 0.010	**0.941** $\pm$ 0.020	**0.898** $\pm$ 0.003	0.053
SSIM	**0.974** $\pm$ 0.003	**0.882** $\pm$ 0.007	0.827 $\pm$ 0.013	0.940 $\pm$ 0.015	0.893 $\pm$ 0.005	0.053
HaarPsi	0.966 $\pm$ 0.004	0.850 $\pm$ 0.020	0.827 $\pm$ 0.062	0.907 $\pm$ 0.017	0.878 $\pm$ 0.009	0.055

TABLE 4.5: Scene by scene SROCC metric for *Reference-based* methods on *unknown content*. The ResNet-18 backbone is used.

4.5.2 *No-Reference* Experiments

We now compare different methods in the case of *no-Reference* setting. As detailed in Section 4.4, we focus on the *unknown content* scenario because of its practical relevance. The

Model	LCC	SROCC	KROCC
BRISQUE [92]	0.151	0.221	0.156
NIQE [93]	0.111	0.061	0.034
IL-NIQE [148]	0.622	0.499	0.361
CNNIQA [65]	0.740 ± 0.058	0.648 ± 0.044	0.486 ± 0.042
WaDIQaM-NR [14]	0.492 ± 0.021	0.421 ± 0.027	0.313 ± 0.021
DBCNN [153]	0.827 ± 0.024	0.746 ± 0.046	0.576 ± 0.051
PaQ-2-PiQ [142]	0.861 ± 0.040	0.823 ± 0.078	0.660 ± 0.084
MUSIQ [66]	0.851 ± 0.058	0.799 ± 0.120	0.636 ± 0.119
VGG-16	0.864 ± 0.009	0.832 ± 0.016	0.669 ± 0.015
ResNet-18	0.870 ± 0.007	0.851 ± 0.007	0.694 ± 0.008
ResNet-50	**0.875** ± 0.009	0.851 ± 0.007	0.690 ± 0.004
MobileNet-V2	0.872 ± 0.005	**0.858** ± 0.012	**0.696** ± 0.014

TABLE 4.6: *No-Reference* experiments evaluated on *unknown content*.

performances of a vanilla ResNet-18 on *known content* can be found in Table 4.2 and is discussed in Section 4.5.1 In this comparison, we include several IQA methods described in Sec.4.2. Since NIQE and IL-NIQE do not require any training, they can be directly evaluated on our dataset. In the case of BRISQUE, we report the performance obtained without retraining the SVR model. Other methods were finetuned on our training dataset starting from the publicly available pre-trained weights. In the particular case of WaDIQaM-NR, we train the network from scratch since pre-trained weights are not available.

Quantitative results are reported in Table 4.6. We observe that classical methods such as BRISQUE and NIQE are not effective on our dataset. BRISQUE, NIQE performances are extremely poor (about 0.1 of LCC). IL-NIQE improvement shows a higher correlation but remains far from satisfactory. The very simple one-layer deep CNNIQA shows performance above these methods, while WaDIQaM-NR performances are under expectations. This can be explained by the limited size of our dataset compared to the dataset used in their original paper.

DBCNN and PaQ-2-PiQ performances are close to non-IQA-specific backbones. Finally, MUSIQ reaches performance slightly lower than deep CNN-based methods(-0.020 of LCC

compared to best performing method). This can be probably explained by the relatively small size of our dataset compared to the transformer architecture requirements. Regarding the deep CNN approaches, except VGG-16 backbone, every tested backbone provides a similar level of performance.

4.5.3 Experimental conclusions and recommendations

We have performed several empirical evaluations of the introduced practical settings and scenarios. From the analysis of these experiments, we are able to provide practical recommendations:

- Siamese architectures with vanilla backbone networks can outperform literature approaches initially designed for full-reference IQA.
- In the *known content* scenario, while a simple vanilla ResNet-18 will provide excellent results, performances can be improved using a siamese network if a reference image is available. In this case, we recommend either the HaarPsi, SSIM or correlation aggregation to combine the features coming from each network branch.
- For the more challenging scenario where the evaluated content is unknown, we recommend either a feature-map-wise correlation or SSIM if a reference image is available.
- We recommend applying the aggregation layer at different scales following the feature pyramid strategy.
- Regarding no-reference CQA, we also observe that recent methods from the IQA literature underperform vanilla CNNs.

4.6 Conclusion

In this chapter, we study for the first time the problem of Camera Quality Assessment from real-world images with natural distortions. To this aim, we collected a large dataset composed of images taken with multiple cameras. The proposed dataset can be distinguished from existing datasets since every scene is shot with multiple devices allowing robust annotation and content-specific evaluation. We have introduced different practical settings and scenarios for learning-based camera-quality assessment. From this empirical evaluation, we provided several practical recommendations. We believe that our novel dataset for CQA and the methods proposed in this chapter will be helpful to design more effective algorithms specific to camera quality assessment. Therefore, this dataset is used in the next chapter (Chapter 5) for a use case on authentic camera distortions for a method that uses multiple samples from the same scenes.

Chapter 5

Test your samples jointly: Pseudo-reference for image quality evaluation

5.1 Introduction

As seen in Chapters 2 and 4, Image quality assessment (IQA) can be addressed in two different settings: no-reference IQA, which consists in estimating the quality of an image without additional information, and full-reference IQA, where we assume that we have at our disposal a high-quality or pristine image that is used to predict the quality of a degraded image. In this chapter, we explore a variant of no-reference IQA where we assume that at test time the goal is to estimate the quality score of different images depicting the same content. In this setting, we can take advantage of the multiple distorted images by modeling the variability over the different test samples. This allows us to provide content context to the evaluated samples as would a reference, without requiring a reference for the scene.

This new setting is motivated by several use cases. It is especially relevant to the case of image quality assessment for camera evaluation. In this task, different cameras are usually compared on the same content [127] [128]. The reference image is not available when it comes to an evaluation in in-the-wild conditions on natural content. Another example is the

case of image enhancement where the reference image is unknown and more accurate quality evaluation algorithms could lead to better enhancement.

To address this new setting, we introduce a new network architecture and its corresponding training strategy. Our architecture allows information exchange across samples of the input batch. More precisely, we train our network to compute a pseudo-reference that describes the evaluated scene. At test time, our method, Pseudo-Reference for Image Quality (PRIQ), is given registered samples of a new scene. The pseudo-reference is predicted by a sub-network that combines features from the different test samples. We perform extensive ablations experiments and compare the performances of the proposed method with state-of-the-art approaches on three different datasets.

5.2 Related Work

We briefly review in this section the methods tested in our state-of-the-art comparison and methods using some kind of pseudo-reference. For a more detailed state-of-the-art, refer to Chapter 2. While earlier no-reference methods used handcrafted features such as Natural Scene Statistics [92, 148], or handbook of features [140], the best-performing methods in image quality assessment are nowadays learning-based. Even though standard vision architectures provide solid baselines, several domain-specific methods have been proposed: Bosse *et al.*propose to equip their convolutional neural network with a patch weighting estimator. Zeng *et al.*[144] use annotations as a Gaussian distribution around the ground truth instead of a single value score. The DBCNN architecture [153] proposed by Zhang *et al.*is composed of two different convolutional models: the first one is trained on the classification of synthetic distortions while the second one is a pre-trained VGG [117] on ImageNet [31], representing perceptual features of natural images. Su *et al.*[118] proposed to disambiguate content features and quality features with the help of a content understanding hypernetwork. Techniques from the computer vision literature have been applied to the image quality assessment task: Meta-learning techniques have been applied by Zhu *et al.*[157] in order to improve the performances to unknown distortions. CONTRIQUE [84], proposed by by Madhusudana *et*

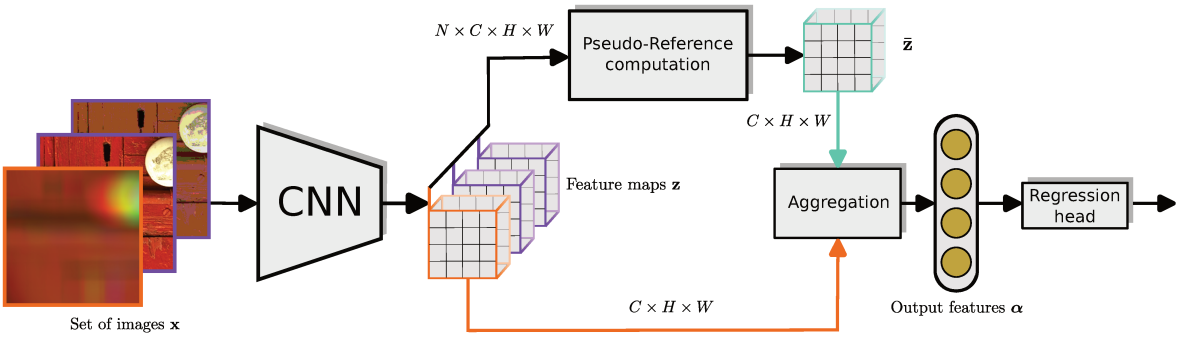


FIGURE 5.1: Illustration of the proposed model. The evaluation pipeline of the evaluated image is highlighted in red. In practice, this procedure is applied in parallel for all the images in the batch. After a transformation of the set of images into feature maps, all these feature maps are used to compute the pseudo-reference, which is then aggregated with the evaluated image's feature maps to produce the output features used to predict quality.

*al.*apply contrastive learning techniques, training their network to produce relevant features for image quality on a large unlabeled dataset. Finally, with the rise of transformers for vision tasks, Ke *et al.*[66] and Golestaneh *et al.*[44] proposed novel architectures to tackle the no-reference image quality problem. Despite remarkable progresses due to more advanced architectures and training procedures, all these methods treat the test samples independently while jointly modeling samples can lead to better predictions.

In no-reference IQA literature, very few works employ a pseudo-reference to devise a formulation closer to full-reference IQA methods. In particular, Lin *et al.*[80] introduces an image generator that estimates a "hallucinated reference" from a distorted image. In the same vein, Zheng *et al.*[155] learns a network that projects the reference and any distorted image to the same feature representation. At test time, another distorted image is provided in order to estimate the pseudo-reference features used to evaluate the target image. Although these two methods are based on pseudo-reference estimation, our approach fundamentally differs from these works since we do not need reference images at training time.

5.3 Method

In this chapter, we assume that we have at our disposal D different scenes $\{\mathcal{S}_1, ..., \mathcal{S}_D\}$ with each scene containing N_d samples of different qualities depicting the same content. Our goal is to estimate image quality scores on novel scenes. While existing methods individually test each sample, our method jointly predicts the score of several samples. We design our approach to enable a joint prediction at test time: while each image will be mapped to a different quality score, its prediction depends on all images in the set. The underlying idea is that, since the samples correspond to different unknown distortions, they can provide contextual information and allow the network to differentiate image features related to content from features related to quality.

To achieve this objective, we introduce the architecture illustrated in Fig. 5.1 where predictions depend on other images in the input set. More specifically, after projecting the images into feature maps, these images are combined into a single pseudo-reference, of the same shape as the feature maps of one image. For each image, these pseudo-reference feature maps are aggregated with each image of the set with using methods inspired by the full-reference literature[40, 128, 130]. These features are then fed into a single-layer regression head which predicts the image quality scores.

5.3.1 Pseudo-Reference computation

Let us assume a convolutional backbone network and a set of N images $\mathbf{x}_i^d$, $i \in [1; N]$, $d \in [1; D]$ from the d^{th} scene. At training time, this set is randomly sampled from the N_d training images of the corresponding scene. At test time, the user can use all the images available. At an intermediary layer of the convolutional network, this set is represented by a tensor $\mathbf{z} = (z_{ickl})$, of dimension $N \times C \times H \times W$, where C denotes the number of channels per image at this specific network layer, and H and W are the height and width of the tensor. We aim to compute a tensor $\bar{\mathbf{z}}$, of dimension $C \times H \times W$, acting as a pseudo-reference for this set of images.

We propose to estimate the pseudo reference with a weighted mean from the feature maps of each image in the set. In this aim, we compute the weights $\mathbf{w}$ measuring the relevance the network should give to each image. More precisely, we perform this computation for every image and location to obtain a pseudo-reference that fully benefits from all the images in the set. Therefore, $\bar{\mathbf{z}}$ is given by:

$$\bar{\mathbf{z}} = \sum_{i=1}^{N} \mathbf{w}_i \odot \mathbf{z}_i, \tag{5.1}$$

with $\odot$ being the element-wise multiplication in the $H \times W$ dimensions. The weights $\mathbf{w}$ are predicted by a sub-network with the feature maps $\mathbf{z}$ as inputs through an attention mechanism. More precisely, they are computed with a one-by-one convolutional layer with one kernel in order to output one channel. The weights need to sum to one over the set dimension to implement the weighted mean operation. Therefore, we follow common practices in attention models and employ the softmax activation:

$$\mathbf{w} = \mathrm{Softmax}(\mathrm{Conv}(\mathbf{z})) \tag{5.2}$$

Note that this pseudo-reference is computed once with every image in the set, and the same pseudo-reference is used to evaluate every image in the set.

5.3.2 Aggregation

After describing the pseudo-reference computation, we can now detail the aggregation scheme that we employ to compare every input image to the pseudo-reference. The aggregation layer receives as inputs the features maps $\mathbf{z}_i$ for an image i in the set and the pseudo-reference feature maps $\bar{\mathbf{z}}$ and outputs a feature vector for each input image people. Inspired by the full-reference literature [40, 128, 130], we chose to apply the channel-wise Structural Similarity Index Measure (SSIM).

To compare the two tensors, we treat separately each channel and consider the two feature maps as C samples of two random variables. Assuming that $\mathbf{z}_{ic}$ and $\bar{\mathbf{z}}_c \in \mathbb{R}^{H \times W}$ respectively denote the values in $\mathbf{z}_i$ and $\bar{\mathbf{z}}$, the output vector $\boldsymbol{\alpha} = \left(\alpha_c\right)_{1 \leq c \leq C}$ for the i-th sample of the

SSIM aggregation layer is given by:

$$\alpha_c = \mathrm{SSIM}(\mathbf{z}_{ic}, \bar{\mathbf{z}}_c) \tag{5.3}$$

5.3.3 Feature pyramid and prediction

Selecting the right scale for comparing the images with the pseudo-reference is not straightforward. We propose to employ a feature pyramid strategy [81, 66, 128] which provides a set of features that are used to evaluate image quality at different scales. The pseudo-reference computation and the aggregation are applied to the output of 5 different intermediate layers of the backbone network at different resolutions. The extracted features at each resolution are then concatenated and fed to a regression head for the image quality prediction.

5.4 Experiments

5.4.1 Evaluation Protocol

Datasets. To evaluate the proposed method, we perform experiments on three datasets:

- TID2013 [98] consists of 25 different contents and 24 distortions with 5 levels each, for a total of 3000 images
- KADID10k [79] consists of 81 different contents and 25 distortions with 5 levels each, for a total of 10125 images
- CARCO dataset [128] presents distortions from numerous camera devices. It differs from other in-the-wild datasets which do not present the same content with varying qualities [57, 41] or lack of region selection and registration [38].

The images of these three datasets are registered. Registration is exact by design for TID2013 and KADID10k while the input images were approximately registered by a prepossessing algorithm for CARCO.

Metrics. Regarding evaluation metrics, we follow previous works [92, 93] and use the Linear Correlation Coefficient (LCC) between the annotations and predictions. Additionally, we report the Spearman Rank-Order Correlation Coefficient (*SROCC*) defined as the linear correlation coefficient of the ranks of predictions and annotations. We report the median of these metrics across the different runs as in [44, 84].

Protocol. In the TID2013 and KADID10k experiments, we select randomly 80% of the datasets' scenes for training and 20% for testing. This protocol ensures there is no content overlap between the training and testing sets. We repeat this process five times and we report the median. The split is fixed over multiple experiments, ensuring different experiments are compared with the same split configuration. For the experiments on CARCO, the dataset size allows us to effectively test the ten scenes independently as the test scene in a ten-fold cross-validation. This process is also repeated five times for more robust results and we also report the median.

At test-time, we simulate different scenarios for a user based on the number of images T available. The number of images T is analogous to the set size N at training time. To simulate these conditions, we randomly split our test dataset into multiple sets of size T. To ensure a fair comparison, we employ the same random sets for every method. The reported correlation metrics are computed for all the samples of the test dataset jointly. Note that set size T at test-time does not correspond to the size of our test dataset.

5.4.2 Implementation Details.

We employ a ResNet-18[53] backbone, pre-trained on ImageNet [31]. The chosen five pyramid stages are placed after the initial 7×7 convolution and after each residual block. We use the Huber loss and train for 60 epochs, with a weight decay of 3×10^{-3} every ten epochs. We used a training batch of 30 images, composed of 6 sets of $N = 5$ images per batch. The images are randomly cropped to a $224 \cdot 224$ size and randomly flipped. All the images of a set are augmented in the same manner in order to preserve alignment.

Module	Weights dimension			T	LCC	SROCC
	N	C	$H \times W$			
(i)	✗	✗	✗		0.890	0.850
(ii)	✓	✗	✗		0.873	0.849
(iii)	✓	✓	✗	5	**0.899**	0.847
(iv)	✓	✗	✓		**0.899**	**0.861**
(v)	✓	✓	✓		0.897	0.856
(i)	✗	✗	✗		0.904	0.879
(ii)	✓	✗	✗		0.893	0.872
(iii)	✓	✓	✗	20	0.911	0.870
(iv)	✓	✗	✓		**0.918**	**0.883**
(v)	✓	✓	✓		0.908	0.879
(i)	✗	✗	✗		0.906	0.887
(ii)	✓	✗	✗		0.913	0.883
(iii)	✓	✓	✗	100	0.909	0.881
(iv)	✓	✗	✓		**0.926**	**0.899**
(v)	✓	✓	✓		0.908	0.886

TABLE 5.1: Analysis of performances of various modules producing a pseudo-reference. T represents the set size at test time. The median correlations over 5 runs are reported.

5.4.3 Analysis: Pseudo-Reference Computation.

We now evaluate different approaches to compute the pseudo-reference feature maps. We explore different solutions regarding the dimensionality of the weights used in the weighted average used to estimate the pseudo-reference.

In section 5.3, the pseudo-reference is computed following equation (5.1). In this analysis, we evaluate this design choice and consider the following variants:

(i) A first naive approach is to compute the mean along the set axis: $\bar{\mathbf{z}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i$.

(ii) We consider a slightly more complex solution based on a weighted average: $\bar{\mathbf{z}} = \sum_{i=1}^{N} \mathbf{w}_i \mathbf{z}_i$, with $\mathbf{w} \in [0,1]^N$. The weights are computed through a linear layer: $\mathbf{w} = \mathrm{Softmax}(\mathbf{A}(\mathrm{GAP}(\mathbf{z}) + \boldsymbol{b}))$, with $\mathbf{A}$ and $\boldsymbol{b}$ being the parameters of a linear layer with

an output size of one, and GAP designating the *Global Average Pooling* operation, with a Softmax computed along the set axis.

(iii) We modify the previous approach to allow different weights for channels of the set images. Here, the output size of the linear layer is C, and $\mathbf{w}$ is of size $N \times C$.

(iv) Instead of weighting the channels, we weigh in this module the feature maps' locations. This corresponds to our proposed model as described in Sec. 5.3.

(v) It is also possible to weigh the locations and channels simultaneously. Thus, the convolutional layer from **(iv)** has now C different kernels instead of one and thus outputs feature maps with C channel for each image. In this case, $\mathbf{w}$ is of size $N \times C \times H \times W$.

The results are reported in Table 5.1. We observe that the setting with the weights $\mathbf{w}$ in $\mathbb{R}^{N \times H \times W}$ provides the best result for every value of T (*eg.* 0.013 SROCC increase for 20 samples compared to the second best performing module). Adding the channel dimension to the set of weights generally degrades performances. A simple mean provides good results and is the second or third-best-performing module depending on the set size.

In Fig.5.2, we provide more values for the set size at test time T to compare these five different modules. These experiments confirm our analysis for intermediary values. Even though our model is trained with a set size of 5, we observe that performances are consistently increasing with the set size at test time. We observe rapid improvements from 2 samples to 10 samples, then a steady rise in performance up to 50 samples, and the results at 100 samples show slight amelioration over the results at 50 samples.

5.4.4 Ablation Study

We now validate the positive impact of each component of the proposed approach. We consider several variants of our approach. First, we train a vanilla ResNet-18 on our datasets, with no pseudo-reference method. Then, we test our method without the feature's pyramid. Instead, a single pseudo-reference is estimated after the last convolutional block. Finally, we

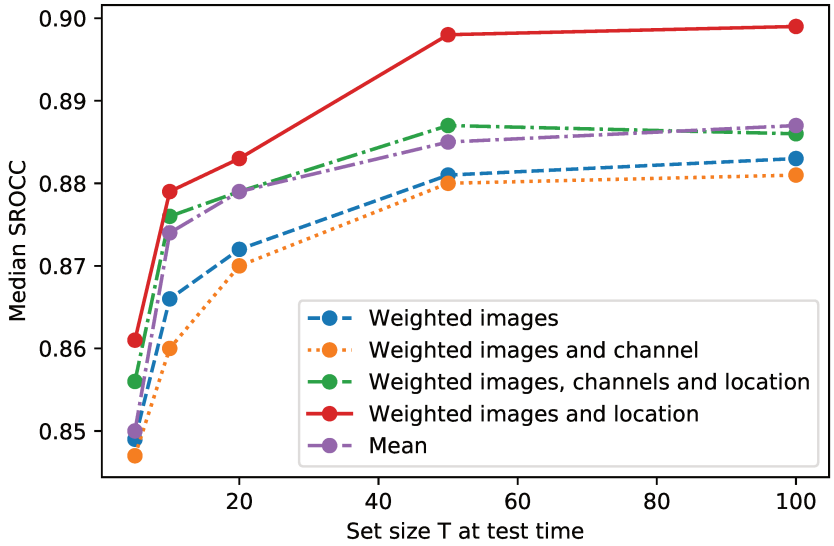


FIGURE 5.2: Impact of the set size T at test time for the different pseudo-reference modes.

replace the SSIM aggregation of our method with a concatenation of features of the image to be evaluated and the pseudo-reference after an average pooling.

P-R	SSIM	Pyr	TID2013		KADID10k		CARCO	
			LCC	SROCC	LCC	SROCC	LCC	SROCC
✗	✗	✗	0.815	0.782	0.825	0.839	0.883	0.853
✓	✓	✗	0.923	0.906	0.934	0.93	0.903	0.864
✓	✗	✓	0.886	0.873	0.883	0.884	0.904	0.873
✓	✓	✓	**0.929**	**0.911**	**0.937**	**0.936**	**0.918**	**0.883**

TABLE 5.2: Ablation study on the *CARCO* dataset: architectural design. We set $T = 20$ for this comparison. The median correlations over 5 runs are reported.

For all three datasets, we found that our method vastly outperforms a Resnet-18 trained regularly. The SSIM aggregation provides a huge improvement over the features' concatenation on the performances in the case of KADID10k and TID2013, while the improvement is still sizeable on CARCO. Finally, even though the feature pyramid provides only small

improvements on the synthetic datasets (TID2013 and KADID10k), the results are still consistently better with the pyramid and its effect on CARCO is important. The clear superiority of SSIM over concatenation on the synthetic dataset can be explained by the fact that SSIM is originally a full-reference metric, based on the assumption that the input images are aligned. This assumption is not strictly respected in the case of the CARCO dataset that contains natural content. In consequence, SSIM may suffer more from this misalignment than concatenation.

5.4.5 Cross-Database Experiment

Cross dataset experiment We test the ResNet-18 baseline and the optimal method deduced from 5.1 and 5.2 in a cross-database experiment: For both methods, we train on CARCO and test on TID2013, and vice-versa. We report the results of the experiment in Tab. 5.3 Even though we observe a drop in performance with respect to the standard setting, we observe that our method outperforms the ResNet-18 baseline in all cases.

Method	Train	Test	LCC	SROCC
ResNet-18	TID	CARCO	0.315	0.262
PRIQ			**0.344**	**0.322**
ResNet-18	CARCO	TID	0.382	0.207
PRIQ			**0.564**	**0.400**

TABLE 5.3: Cross-dataset study of the baseline and the full method for T = 20

5.4.6 Edge-case

We consider the extreme case where all the images but one have low-quality scores and check whether the bad images could obtain high scores because they would be more similar to the pseudo-reference. We obtain the result shown in Fig. 5.3. We observe that the only high-quality image still obtains a high score. It demonstrates the robustness of our approach even in such an extreme case. By looking at the weights maps, we observe that the high-quality image is associated with higher weights than the other images when estimating the

pseudo-reference. It indicates that the attention layers have learned to select high-quality features.

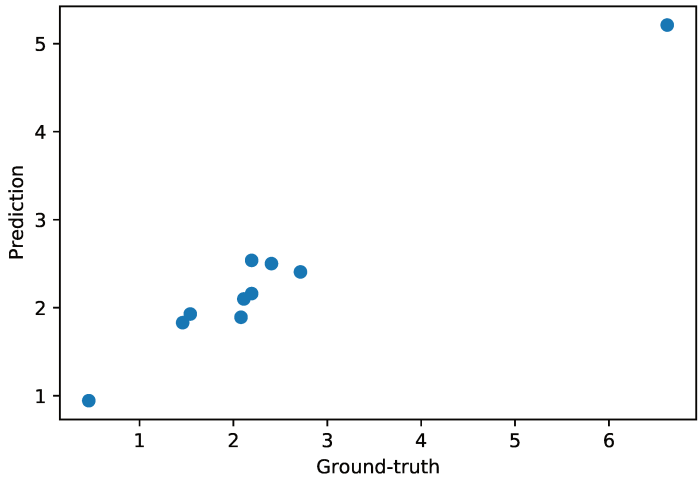


FIGURE 5.3: Predictions obtained with a batch composed of a vast majority of low-quality samples.

5.4.7 Comparison to the State of the Art

The comparison to the state-of-the-art is summarized in Table 5.4. Our method is evaluated here in the optimal setup obtained in ablation studies. Our framework is the first approach that can jointly evaluate multiple images jointly. Therefore, our approach is compared with traditional methods that evaluate images independently. We observe that joint evaluation with our method greatly improves predictions according to every metric with $T \geq 5$ for TID2013, and CARCO, whereas CONTRIQUE [84] achieves great results on KADID10k and achieves performances roughly equal to our method. Notably, we observe that using a set size larger than 20 for our offers little improvement in the case of synthetic datasets. However, the gain in performance might be interesting on authentically distorted images.

We also observe that our method vastly outperforms other approaches on CARCO whose images are authentic camera photographs. This result validates our approach in this particularly relevant setting of camera evaluation. Concerning variants explored in 5.2, the variant without the pyramid still vastly outperforms other methods on the TID dataset, while only

CONTRIQUE is performing similarly on KADID. The variant without the SSIM aggregation is still slightly the best-performing method on TID with the TReS method coming close, while on KADID it is only outperformed by CONTRIQUE. On the CARCO dataset, the no-SSIM performs similarly to the best state-of-the-art method while the no-pyramid variant slightly underperforms the 3 best-performing methods in terms of SROCC.

	TID2013		KADID10k		CARCO	
Method	LCC	SROCC	LCC	SROCC	LCC	SROCC
BRISQUE [92]	0.571	0.626	0.567	0.528	0.151	0.221
IL-NIQE [148]	0.648	0.521	0.558	0.534	0.622	0.499
ResNet-18 [53]	0.815	0.782	0.825	0.839	0.883	0.853
PQR [144]	0.798	0.740	-	-	-	-
WaDIQaM [14]	0.855	0.835	0.752	0.739	-	-
DBCNN [153]	0.865	0.816	0.856	0.851	0.827	0.746
Meta-IQA [157]	0.868	0.856	0.775	0.762	-	-
HyperIQA [118]	0.858	0.840	0.845	0.852	0.889	0.881
TReS [44]	0.883	0.863	0.858	0.859	0.904	0.869
CONTRIQUE [84]	0.857	0.843	**0.937**	0.934	0.900	0.875
PRIQ, T=2	0.857	0.841	0.874	0.869	0.899	0.861
PRIQ, T=5	0.916	0.894	0.933	0.934	0.911	0.879
PRIQ, T=20	0.929	**0.911**	**0.937**	**0.936**	0.918	0.883
PRIQ, T=100	**0.930**	**0.911**	**0.937**	0.935	**0.926**	**0.899**

TABLE 5.4: Comparison to the state-of-the-art. The median correlations over 5 runs are reported. Some results are borrowed from [44] and [84]

5.5 Conclusion

We introduced a setting where images are jointly evaluated that effectively uses several different images of the same content in order to provide semantic contextual information. The proposed design estimates a pseudo-reference at the feature level and employs a feature pyramid aggregation. We conducted an ablation experiment to determine the optimal pseudo-reference computation module and another ablation to understand the contribution

of each part of our method. We performed extensive evaluations across several image quality datasets to validate the efficiency of the proposed method and found that we achieve competitive or better performances than state-of-the-art no-reference image quality methods whether on synthetic or in-the-wild datasets. We encourage IQA researchers to explore this new setting due to the large possibilities for taking advantage of the multiple inputs and the possible applications.

Chapter 6

Conclusion

In conclusion, this Ph.D. dissertation addresses the limitations of current image quality assessment methods for smartphone cameras and proposes innovative solutions using deep learning systems. With the rapid growth of smartphone camera usage, there is a need for accurate measurement and evaluation of image quality criteria specific to smartphone cameras.

First, our study focuses on a case in a laboratory environment. Our method enables the automatic detection of relevant regions given an annotated image quality attribute. This method allows the use of deep learning architectures when dealing with high-resolution smartphone images for the evaluation of local attributes. This approach outperforms traditional chart-based methods.

Second, a new in-the-wild dataset is created to accurately represent the complex mixture of defects commonly found in smartphone camera images, as well as repeated contents that permit the adaptation of the full-reference method. This dataset enables a comprehensive evaluation of different methods in various practical scenarios, providing valuable guidelines for using deep learning systems in camera quality assessment.

Finally, we designed a method that incorporates the assumption in which we have our disposal several samples from the same scenes to enhance the design of more accurate models. A pseudo-reference is computed from available distorted images, bypassing the need for a high-quality reference. This method addresses a hypothesis rarely explored in research datasets, presenting potential applications beyond camera quality assessment.

Overall, this research contributes to advancing the field of smartphone camera image quality assessment by addressing existing limitations and proposing innovative deep learning-based solutions. The findings have practical implications and can be applied to various camera quality attributes, providing valuable insights for researchers and practitioners working in the field of image quality assessment. For the computer vision community, we recommend paying particular attention to the last method presented, combining several samples of the same scene to compute a reference model of the scene. Applications in object detection or instance segmentations are a possibility, in the case of non-moving cameras. Despite the lack of repeated content in the research dataset, many vision solutions are industrially deployed on fixed optic systems.

We recommend several future works following the different findings. First, a thorough study would be interesting in the context of Chapter 4 to determine the behavior of Reference-based methods under controlled registration mistakes. Second, the attention mechanism in Chapter 5 used to compute the pseudo-reference is a highly effective but rather naïve scheme. It would deserve to be further investigated. Furthermore, applying this idea to the Vision Transformer architecture would certainly allow for further performance gain. Finally, we would like to test this method on object detection problems, provided we access appropriate data in which the same cameras are providing data with the same field of view.

Appendix A

Discriminant Maps

In this first appendix, we present the discriminant maps obtained with the method presented in Chapter 3. For both texture and noise, we display maps from 3 groups of devices corresponding to the whole device range and then separately for high-quality and low-quality devices.

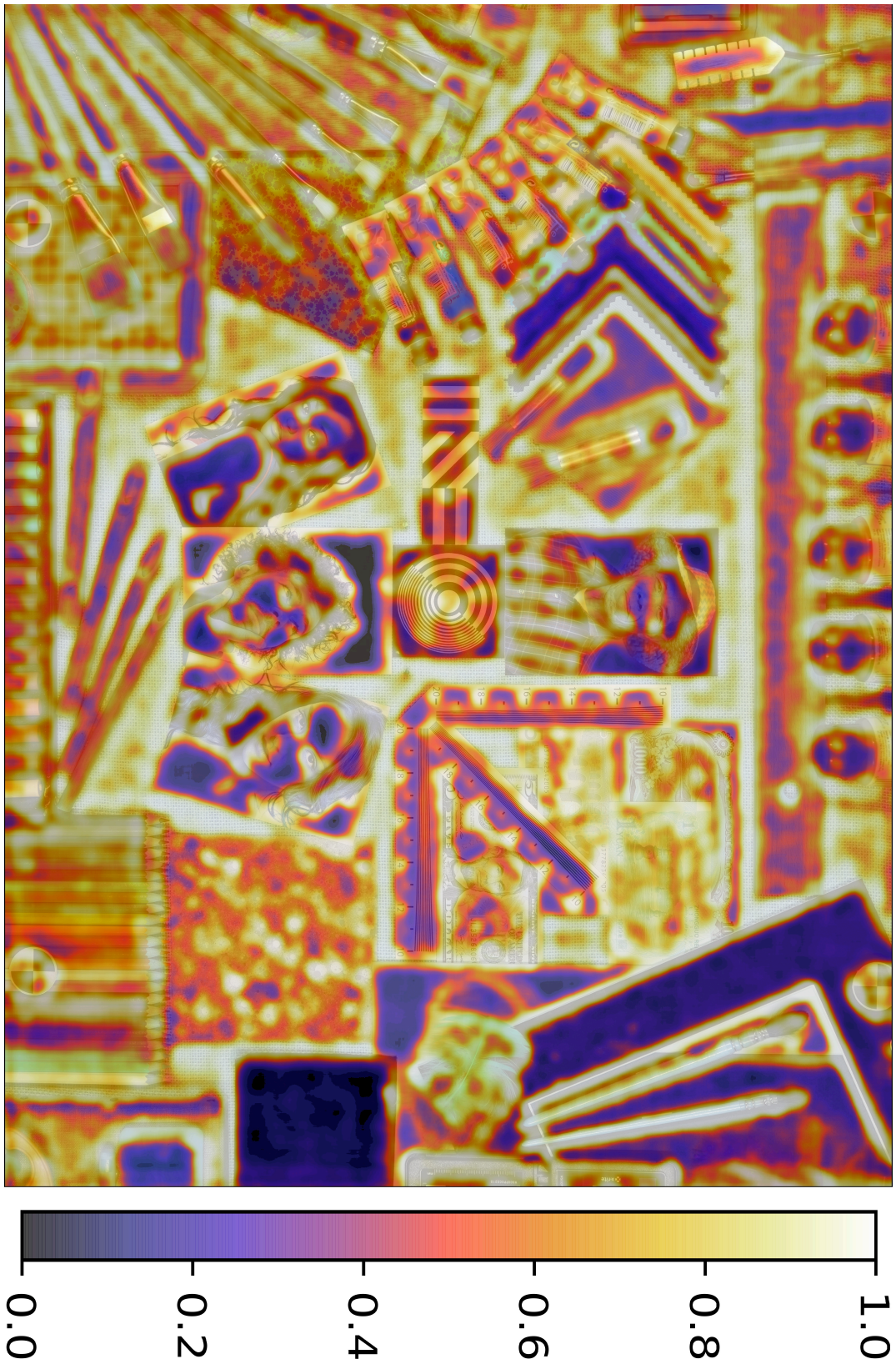


FIGURE A.1: **Whole range** of devices quality discriminant map for **texture** evaluation

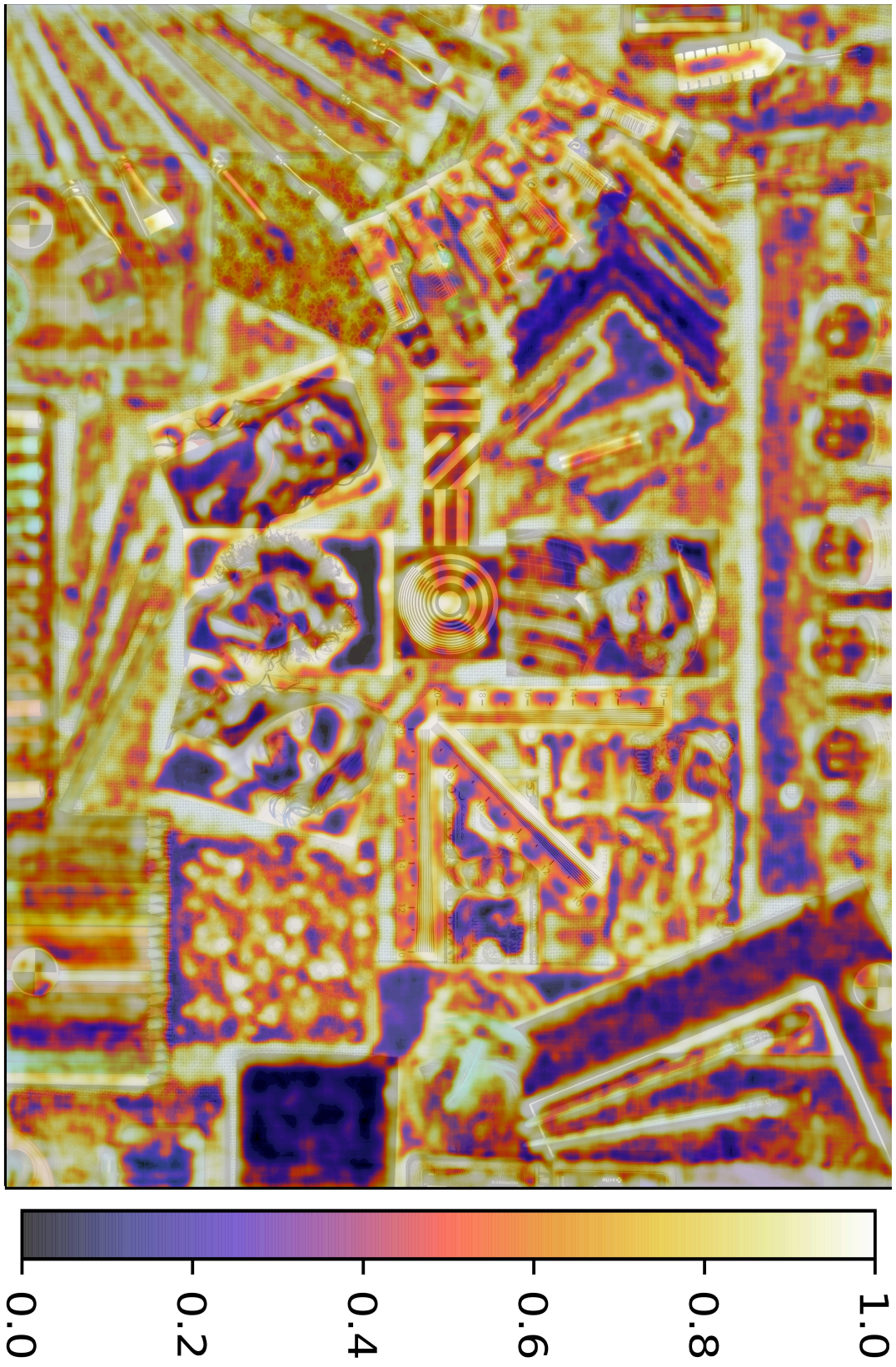


FIGURE A.2: **Low-quality** devices discriminant map for **texture** evaluation

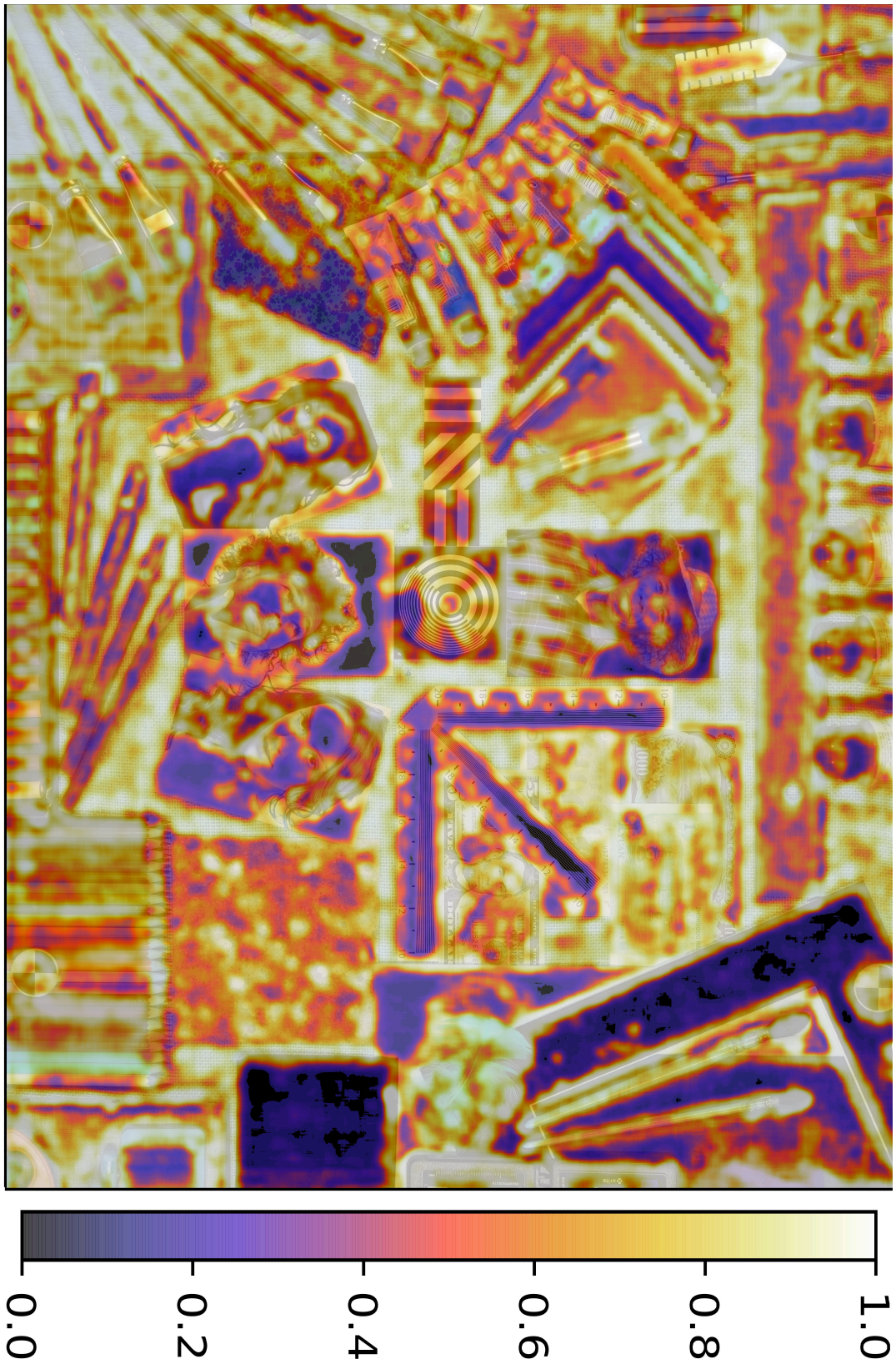


FIGURE A.3: **High-quality** devices discriminant map for **texture** evaluation

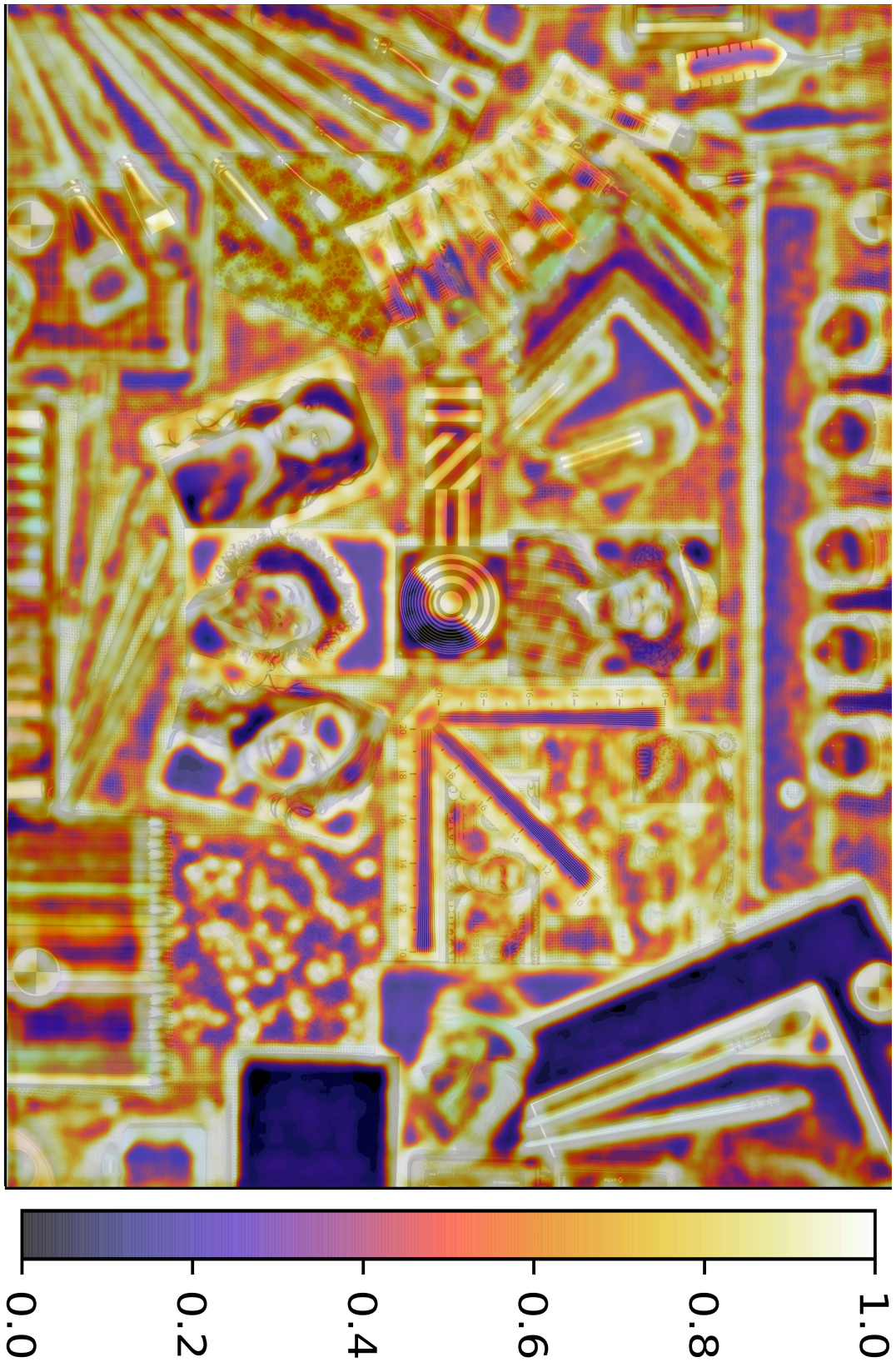


FIGURE A.4: **Whole range** of devices quality discriminant map for **noise** evaluation

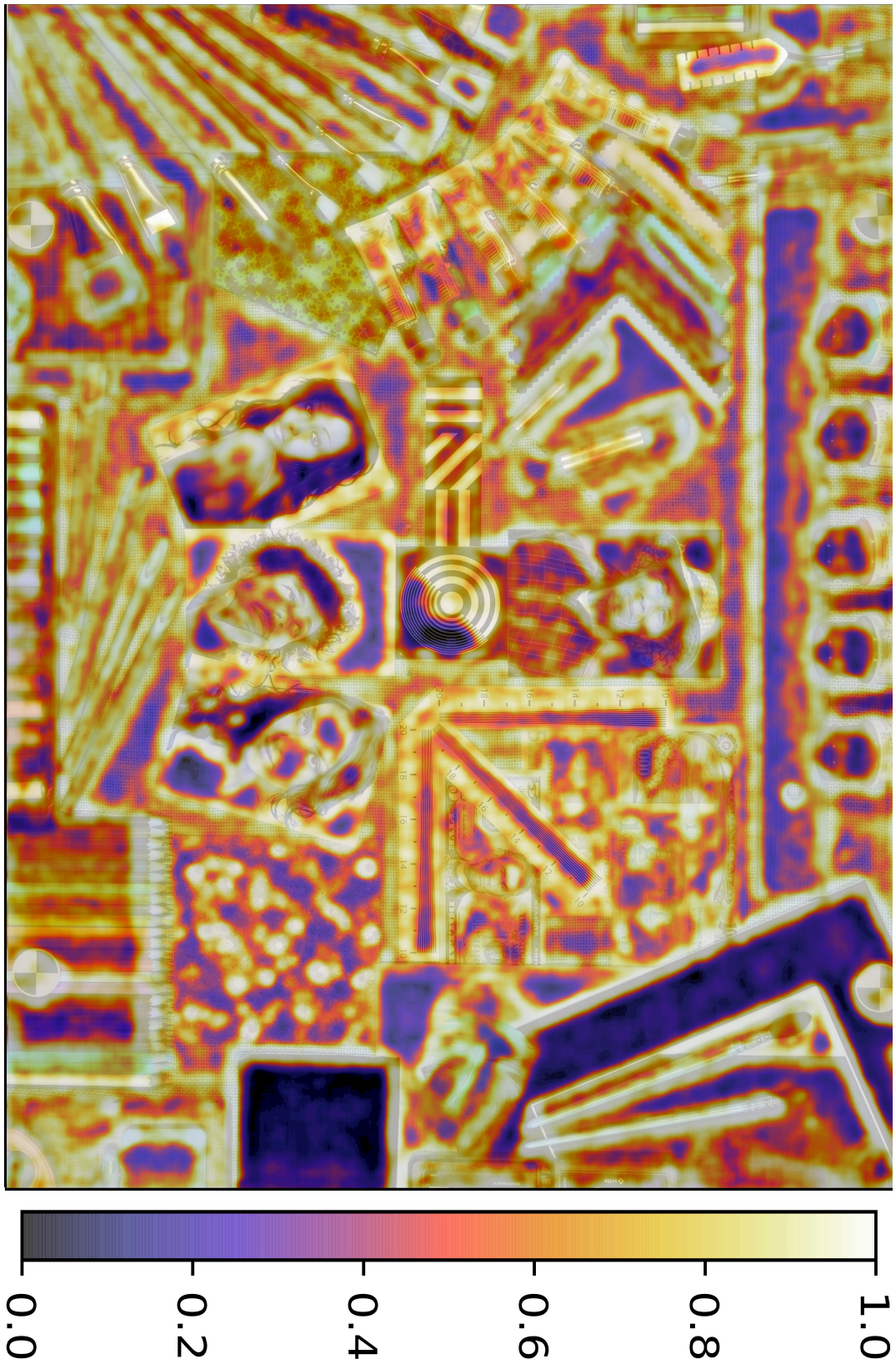


FIGURE A.5: **Low-quality** devices discriminant map for **noise** evaluation

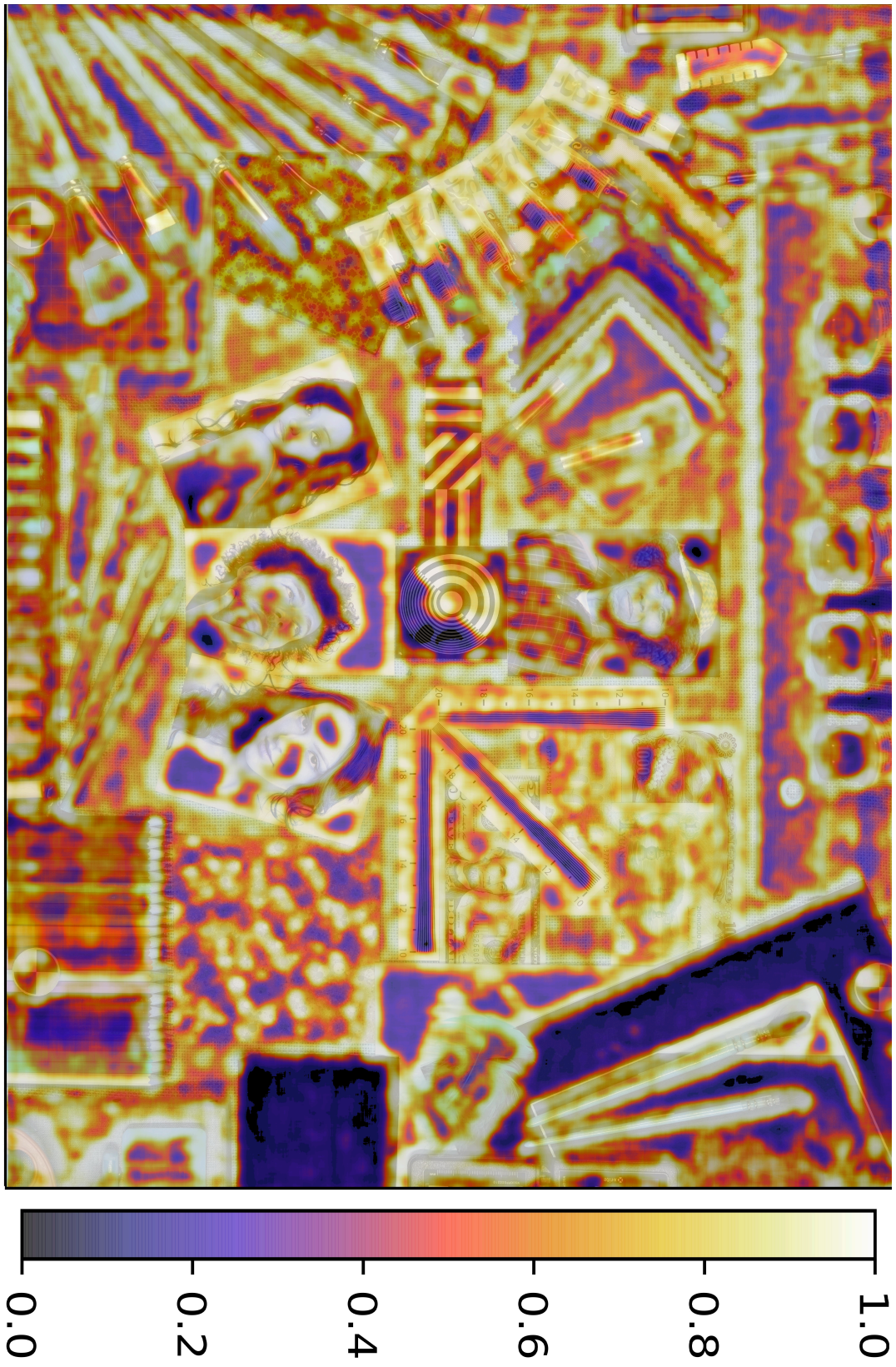


FIGURE A.6: **High-quality** devices discriminant map for **noise** evaluation

Appendix B

CARCO details

In this second appendix we provide more details about the CARCO dataset presented in chapter 4. For each dataset's scene, we provide these four elements:

- The reference image of the scene.
- A visualization of the comparison map from the annotations experiment.
- The scores and the associated uncertainties.
- The distribution density in the form of a histogram.

B.1 Scene #1

FIGURE B.1: Reference texture

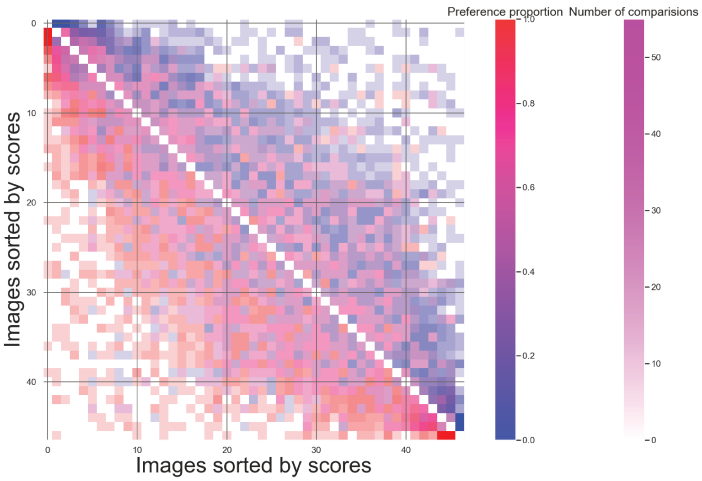


FIGURE B.2: Comparison matrix visualisation

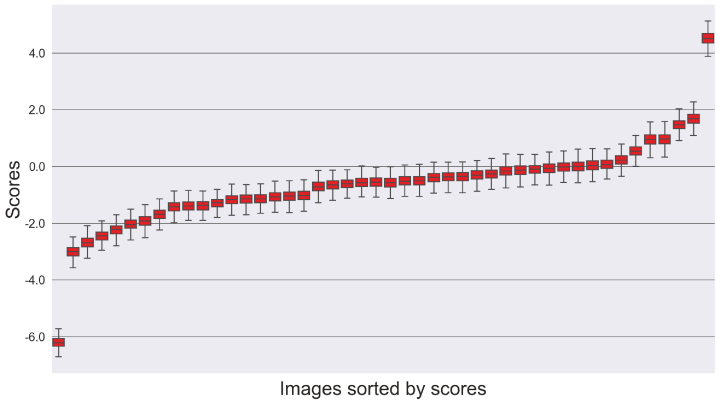


FIGURE B.3: Scores and confidence intervals

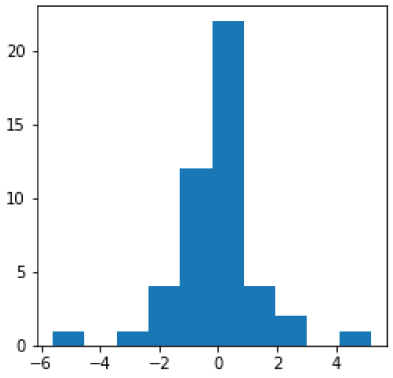


FIGURE B.4: Scores distribution

B.2 Scene #2

FIGURE B.5: Reference texture

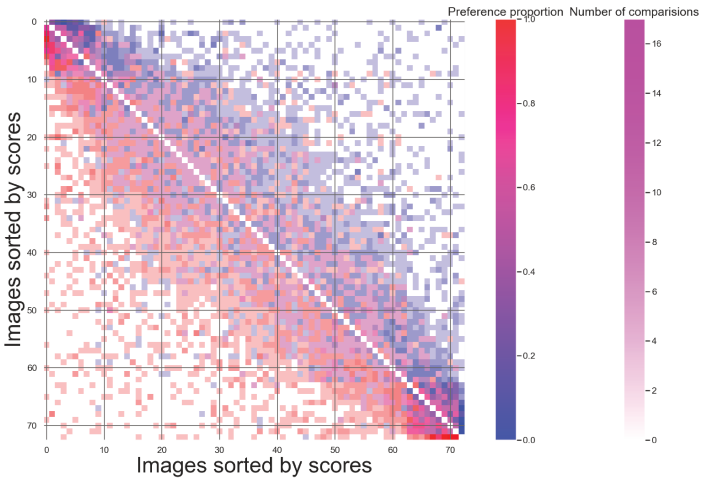


FIGURE B.6: Comparison matrix visualisation

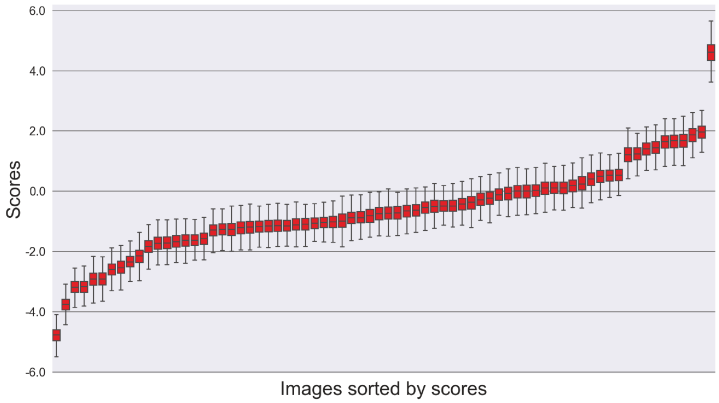


FIGURE B.7: Scores and confidence intervals

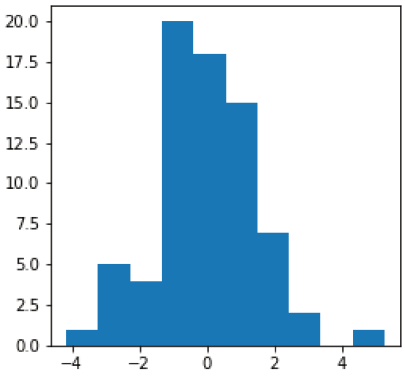


FIGURE B.8: Scores distribution

B.3 Scene #3

FIGURE B.9: Reference texture

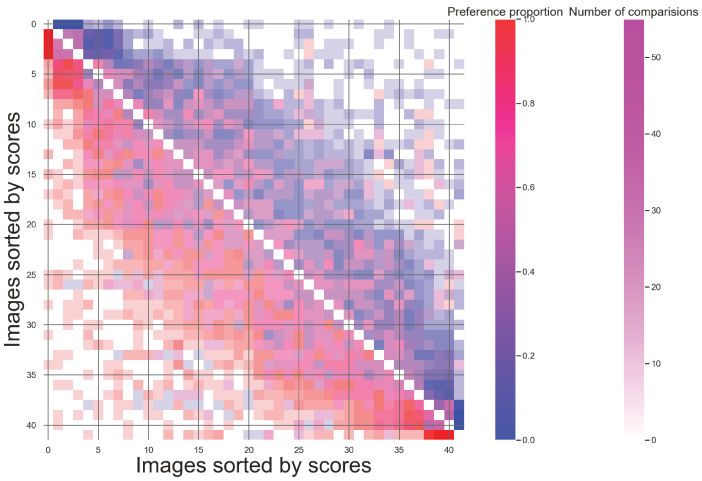


FIGURE B.10: Comparison matrix visualisation

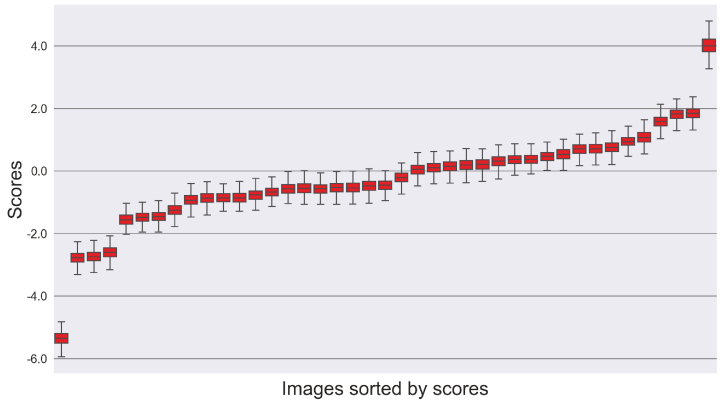


FIGURE B.11: Scores and confidence intervals

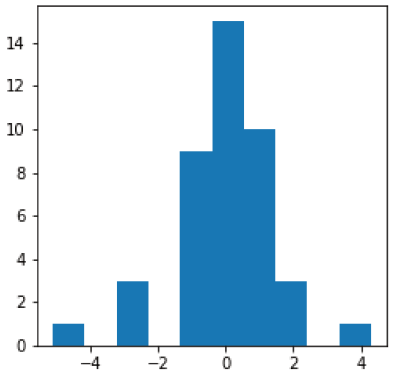


FIGURE B.12: Scores distribution

B.4 Scene #4

FIGURE B.13: Reference texture

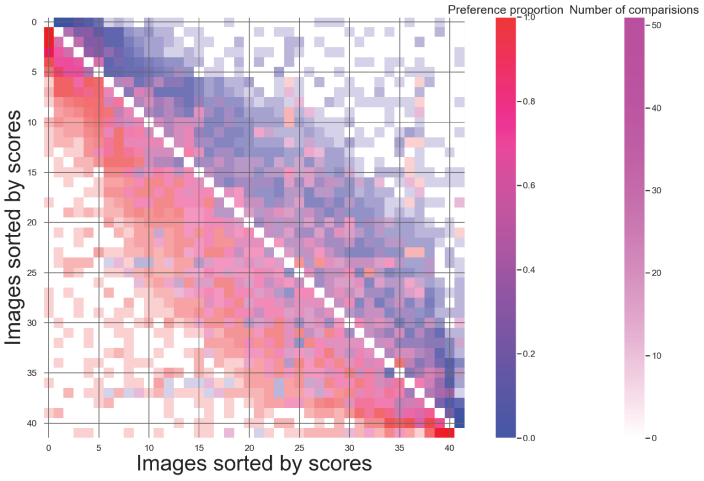


FIGURE B.14: Comparison matrix visualisation

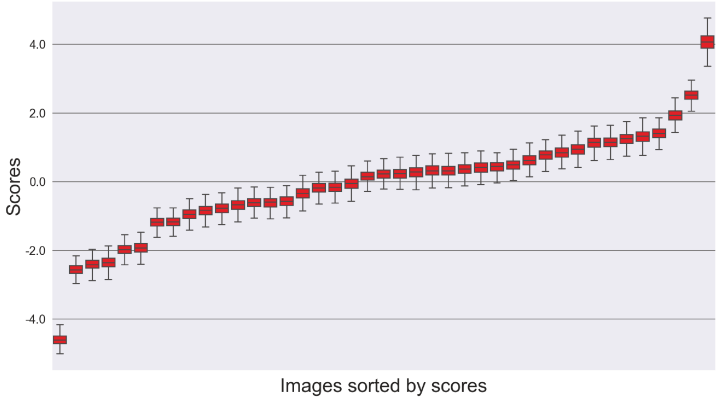


FIGURE B.15: Scores and confidence intervals

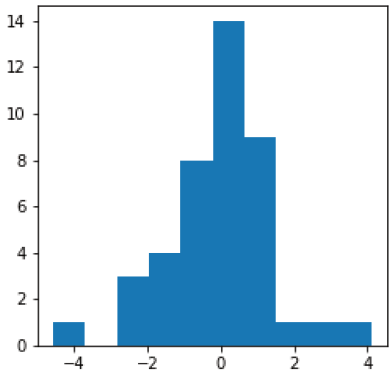


FIGURE B.16: Scores distribution

B.5 Scene #5

FIGURE B.17: Reference texture

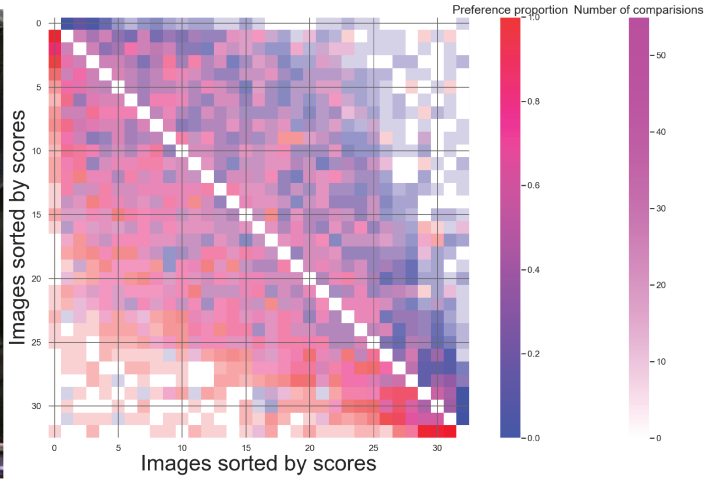


FIGURE B.18: Comparison matrix visualisation

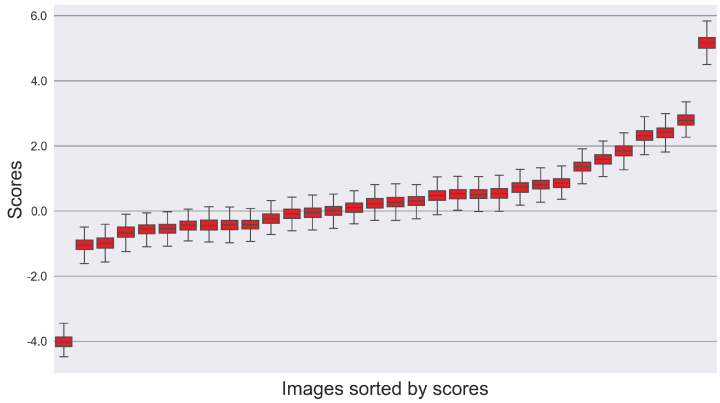


FIGURE B.19: Scores and confidence intervals

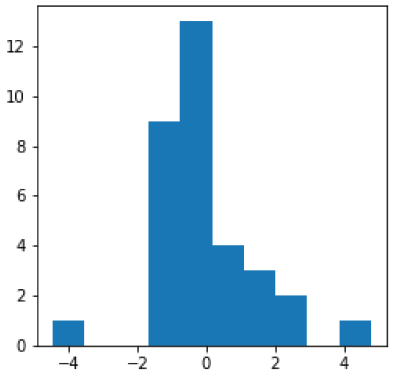


FIGURE B.20: Scores distribution

B.6 Scene #6

FIGURE B.21: Reference texture

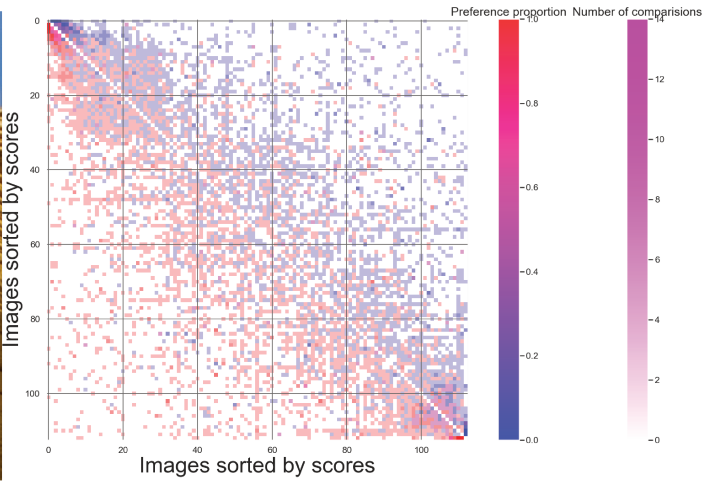


FIGURE B.22: Comparison matrix visualisation

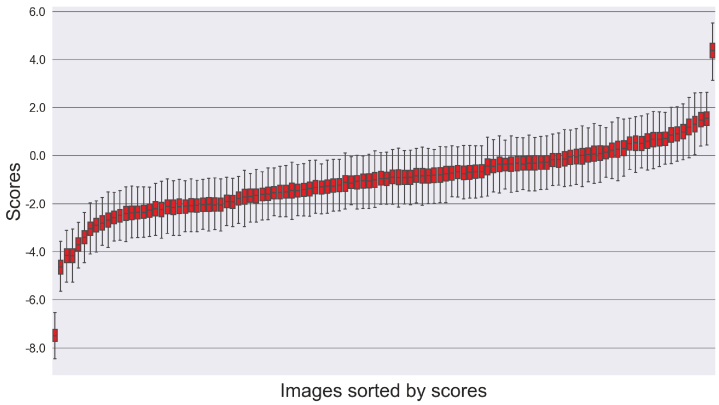


FIGURE B.23: Scores and confidence intervals

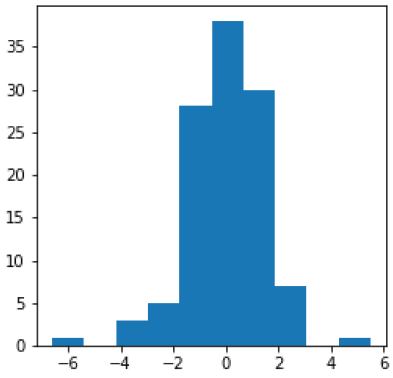


FIGURE B.24: Scores distribution

B.7 Scene #7

FIGURE B.25: Reference texture

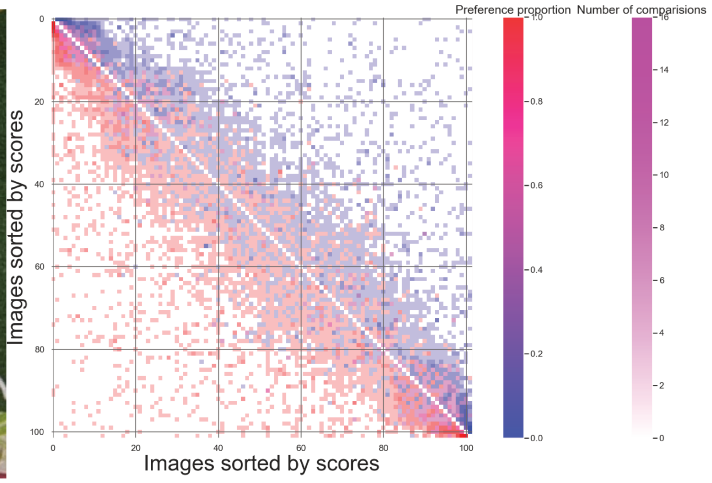


FIGURE B.26: Comparison matrix visualisation

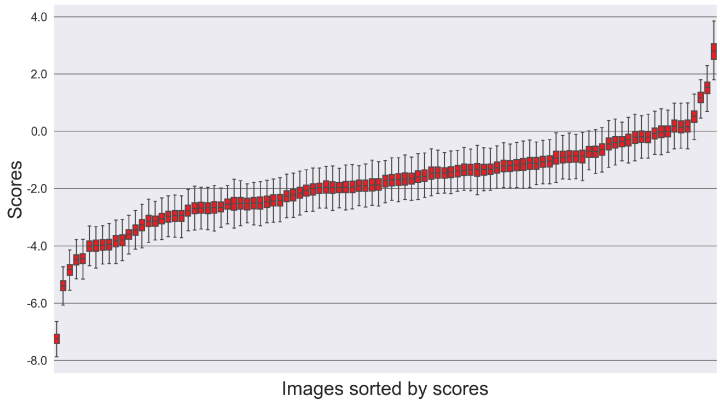


FIGURE B.27: Scores and confidence intervals

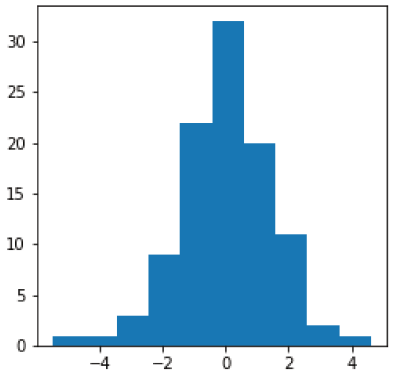


FIGURE B.28: Scores distribution

B.8 Scene #8

FIGURE B.29: Reference texture

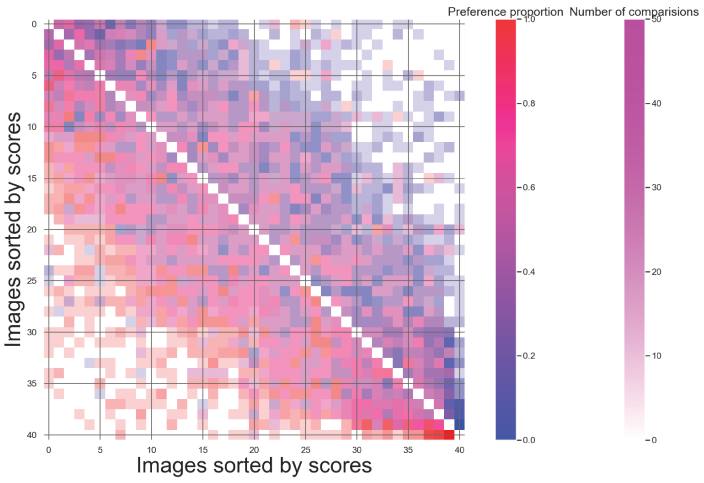


FIGURE B.30: Comparison matrix visualisation

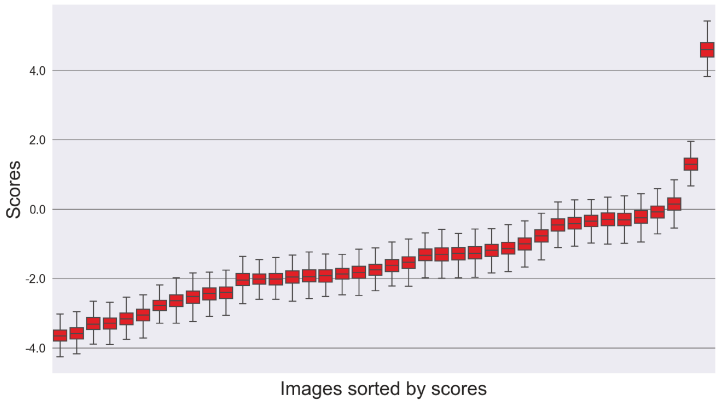


FIGURE B.31: Scores and confidence intervals

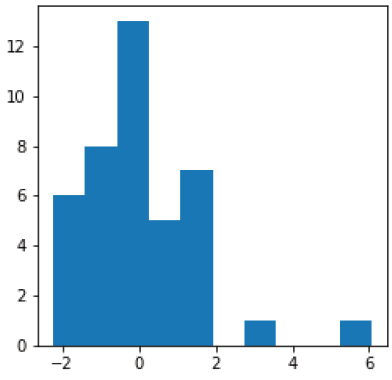


FIGURE B.32: Scores distribution

B.9 Scene #9

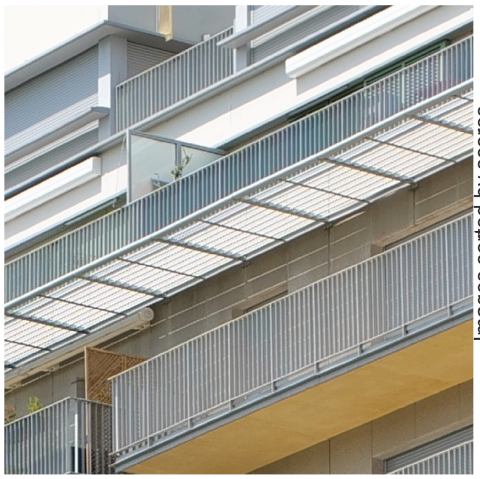


FIGURE B.33: Reference texture

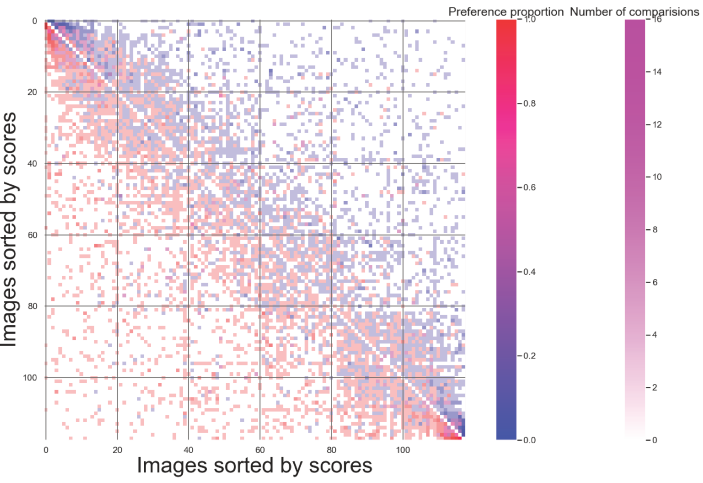


FIGURE B.34: Comparison matrix visualisation

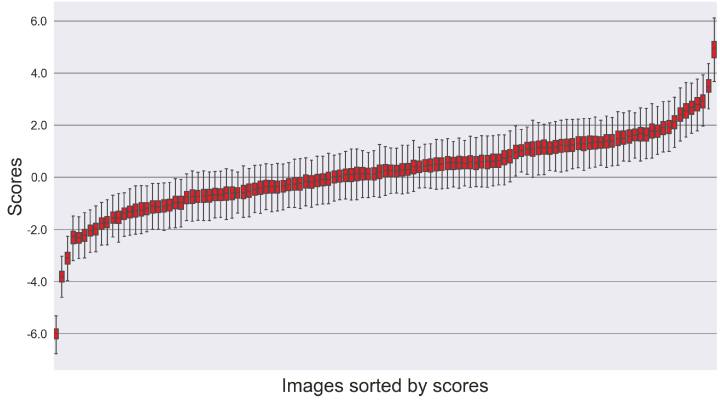


FIGURE B.35: Scores and confidence intervals

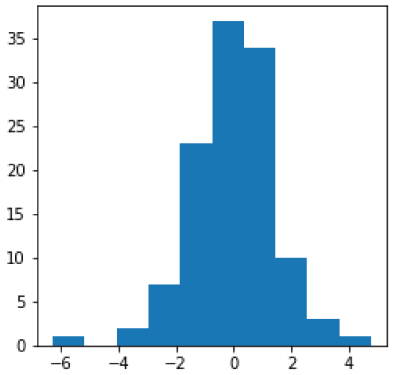


FIGURE B.36: Scores distribution

B.10 Scene #10

FIGURE B.37: Reference texture

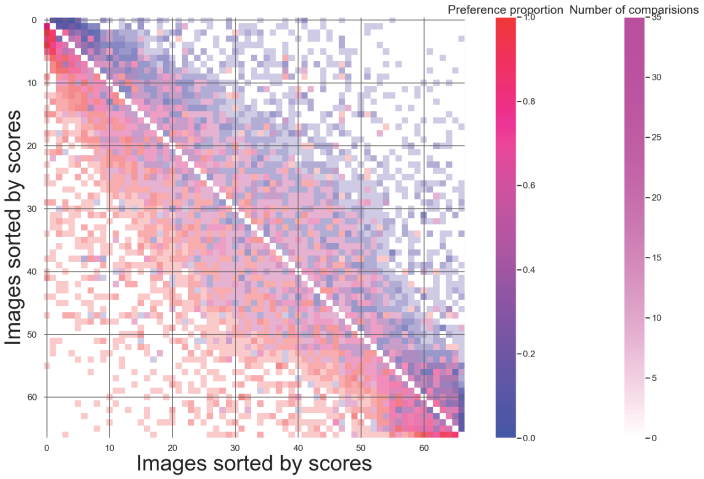


FIGURE B.38: Comparison matrix visualisation

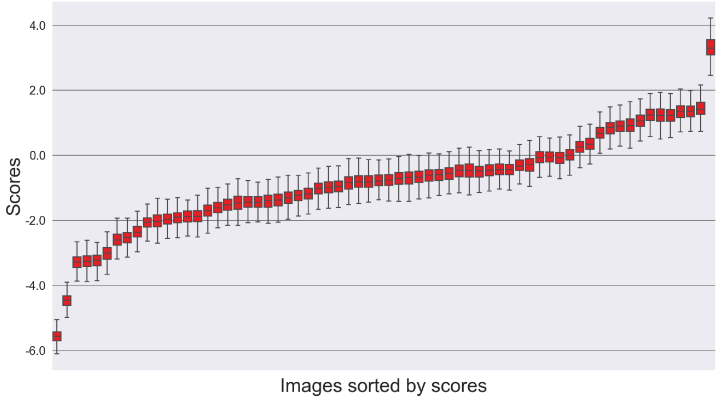


FIGURE B.39: Scores and confidence intervals

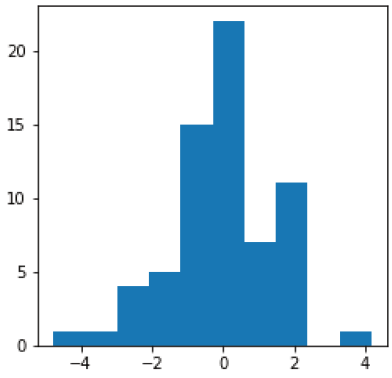


FIGURE B.40: Scores distribution

Bibliography

[1] IEEE standard for camera phone image quality.

[2] Imatest website.

[3] Photography – electronic still picture imaging – noise measurements. Standard, International Organization for Standardization, 2013.

[4] https://www.dxomark.com/smartphones-vs-cameras-closing-the-gap-on-image-quality smartphones vs cameras: Closing the gap on image quality.

[5] AGGARWAL, R., SOUNDERAJAH, V., MARTIN, G., TING, D. S., KARTHIKESALINGAM, A., KING, D., ASHRAFIAN, H., AND DARZI, A. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ digital medicine* (2021).

[6] ALCANTARILLA, P. F., AND SOLUTIONS, T. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE TPAMI* (2011).

[7] ALHASHIM, I., AND WONKA, P. High quality monocular depth estimation via transfer learning. *arXiv preprint* (2018).

[8] BALLÉ, J., LAPARRA, V., AND SIMONCELLI, E. P. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704* (2016).

[9] BELKARFA, S., CHOUKARAH, A. H., AND TWORSKI, M. Automatic noise analysis on still life chart. In *London Imaging Meeting* (2021), vol. 2, Society for Imaging Science and Technology, pp. 101–105.

[10] BHAT, G., DANELLJAN, M., AND TIMOFTE, R. Ntire 2021 challenge on burst super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021).

[11] BOARD, C. A. Ieee standard for camera phone image quality.

[12] BÖCKENHOLT, U. Applications of thurstonian models to ranking data. In *Probability models and statistical analyses for ranking data*. Springer, 1993.

[13] BOREMAN, G. *Modulation Transfer Function in Optical and Electro-optical Systems*. Society of Photo Optical, 2001.

[14] BOSSE, S., MANIRY, D., MÜLLER, K.-R., WIEGAND, T., AND SAMEK, W. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing* (2017).

[15] BOURBON, T., HILLAIRET, C. S., POCHON, B., AND GUICHARD, F. New visual noise measurement on a versatile laboratory setup in hdr conditions for smartphone camera testing. *Electronic Imaging* (2022).

[16] BRANDÃO, T., AND QUELUZ, M. P. No-reference image quality assessment based on dct domain statistics. *Signal processing* (2008).

[17] BRUCE, N. D., AND TSOTSOS, J. K. Saliency, attention, and visual search: An information theoretic approach. *Journal of vision* (2009).

[18] ČADÍK, M., WIMMER, M., NEUMANN, L., AND ARTUSI, A. Image attributes and quality for evaluation of tone mapping operators. In *National Taiwan University* (2006), Citeseer.

[19] CAO, F., GUICHARD, F., AND HORNUNG, H. Measuring texture sharpness of a digital camera. In *Digital Photography V* (2009), International Society for Optics and Photonics.

[20] CARVALHO, M., LE SAUX, B., TROUVÉ-PELOUX, P., ALMANSA, A., AND CHAMPAGNAT, F. On regression losses for deep depth estimation. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (2018), IEEE.

[21] CHEN, J. I.-Z., AND LAI, K.-L. Deep convolution neural network model for credit-card fraud detection and alert. *Journal of Artificial Intelligence* (2021).

[22] CHEN, K.-T., WU, C.-C., CHANG, Y.-C., AND LEI, C.-L. A crowdsourceable qoe evaluation framework for multimedia content. In *Proceedings of the 17th ACM international conference on Multimedia* (2009).

[23] CHEN, T., KORNBLITH, S., NOROUZI, M., AND HINTON, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (2020), PMLR.

[24] CHENG, J., WU, Y., ABDALMAGEED, W., AND NATARAJAN, P. Qatm: Quality-aware template matching for deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019).

[25] CHENG, Z., TAKEUCHI, M., AND KATTO, J. A pre-saliency map based blind image quality assessment via convolutional neural networks. In *2017 IEEE International Symposium on Multimedia (ISM)* (2017), IEEE.

[26] CHEON, M., YOON, S.-J., KANG, B., AND LEE, J. Perceptual image quality assessment with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021).

[27] CONDE, M. V., BURCHI, M., AND TIMOFTE, R. Conformer and blind noisy students for improved image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).

[28] CONG, H., FU, L., ZHANG, R., ZHANG, Y., WANG, H., HE, J., AND GAO, J. Image quality assessment with gradient siamese network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).

[29] CSATÓ, L. Ranking by pairwise comparisons for swiss-system tournaments. *Central European Journal of Operations Research* (2013).

[30] DABOV, K., FOI, A., KATKOVNIK, V., AND EGIAZARIAN, K. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing* (2007).

[31] DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., AND FEI-FEI, L. Image-net: A large-scale hierarchical image database. In *CVPR* (2009).

[32] DING, K., MA, K., WANG, S., AND SIMONCELLI, E. P. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence* (2020).

[33] DONG, C., LOY, C. C., HE, K., AND TANG, X. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence* (2015).

[34] DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBORN, D., ZHAI, X., UNTERTHINER, T., DEHGHANI, M., MINDERER, M., HEIGOLD, G., GELLY, S., ET AL. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[35] ELO, A. E. *The rating of chessplayers, past and present*. Arco Pub., 1978.

[36] ESSER, P., ROMBACH, R., AND OMMER, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 12873–12883.

[37] FACCIOLO, G., PACIANOTTO, G., RENAUDIN, M., VIARD, C., AND GUICHARD, F. Quantitative measurement of contrast, texture, color, and noise for digital photography of high dynamic range scenes. *Electronic Imaging* (2018).

[38] FANG, Y., ZHU, H., ZENG, Y., MA, K., AND WANG, Z. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020).

[39] FINN, C., ABBEEL, P., AND LEVINE, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning* (2017), PMLR.

[40] GAO, F., WANG, Y., LI, P., TAN, M., YU, J., AND ZHU, Y. Deepsim: Deep similarity for image quality assessment. *Elsevier Neurocomputing* (2017).

[41] GHADIYARAM, D., AND BOVIK, A. C. Massive online crowdsourced study of subjective and objective picture quality. *IEEE TIP* (2016).

[42] GIRSHICK, R. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (2015).

[43] GIRSHICK, R., DONAHUE, J., DARRELL, T., AND MALIK, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014).

[44] GOLESTANEH, S. A., DADSETAN, S., AND KITANI, K. M. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *WACV* (2022).

[45] GOUSSEAU, Y., AND ROUEFF, F. Modeling occlusion and scaling in natural images. *Multiscale Modeling & Simulation* (2007).

[46] GU, J., CAI, H., DONG, C., REN, J. S., TIMOFTE, R., GONG, Y., LAO, S., SHI, S., WANG, J., YANG, S., ET AL. Ntire 2022 challenge on perceptual image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022).

[47] GU, J., SHEN, Y., AND ZHOU, B. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020).

[48] GU, S., ZHANG, L., ZUO, W., AND FENG, X. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014).

[49] HAAR, A. *Zur theorie der orthogonalen funktionensysteme*. Georg-August-Universitat, Gottingen., 1909.

[50] HASINOFF, S. W., SHARLET, D., GEISS, R., ADAMS, A., BARRON, J. T., KAINZ, F., CHEN, J., AND LEVOY, M. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM TOG* (2016).

[51] HAUSER, W., NEVEU, B., JOURDAIN, J.-B., VIARD, C., AND GUICHARD, F. Image quality benchmark of computational bokeh. *Electronic Imaging* (2018).

[52] HE, K., GKIOXARI, G., DOLLÁR, P., AND GIRSHICK, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (2017).

[53] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. In *IEEE TPAMI* (2016).

[54] HERBRICH, R., MINKA, T., AND GRAEPEL, T. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems* (2006).

[55] HOERL, A. E., AND KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* (1970).

[56] HOSU, V., LIN, H., AND SAUPE, D. Expertise screening in crowdsourcing image quality. In *2018 Tenth international conference on quality of multimedia experience (QoMEX)* (2018), IEEE.

[57] HOSU, V., LIN, H., SZIRANYI, T., AND SAUPE, D. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE TIP* (2020).

[58] HOU, X., HAREL, J., AND KOCH, C. Image signature: Highlighting sparse salient regions. *IEEE transactions on pattern analysis and machine intelligence* (2011).

[59] HOWELL, D. C. *Statistical methods for psychology*. Cengage Learning, 2012.

[60] JAIN, V., AND SEUNG, S. Natural image denoising with convolutional networks. *Advances in neural information processing systems* (2008).

[61] JIANG, H., AND LEARNED-MILLER, E. Face detection with the faster r-cnn. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)* (2017), IEEE.

[62] JINJIN, G., HAOMING, C., HAOYU, C., XIAOXING, Y., REN, J. S., AND CHAO, D. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16* (2020), Springer.

[63] JO, Y., YANG, S., AND KIM, S. J. Investigating loss functions for extreme super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (2020).

[64] JUMPER, J., EVANS, R., PRITZEL, A., GREEN, T., FIGURNOV, M., RONNEBERGER, O., TUNYASUVUNAKOOL, K., BATES, R., ŽÍDEK, A., POTAPENKO, A., ET AL. Highly accurate protein structure prediction with alphafold. *Nature* (2021).

[65] KANG, L., YE, P., LI, Y., AND DOERMANN, D. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014).

[66] KE, J., WANG, Q., WANG, Y., MILANFAR, P., AND YANG, F. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021).

[67] KINGMA, D., AND BA, J. Adam: A method for stochastic optimization. *ICLR* (2015).

[68] KIRK, L., HERZER, P., ARTMANN, U., AND KUNZ, D. Description of texture loss using the dead leaves target: current issues and a new intrinsic approach. In *Digital Photography X* (2014), International Society for Optics and Photonics.

[69] KOCH, G., ZEMEL, R., AND SALAKHUTDINOV, R. Siamese neural networks for one-shot image recognition. In *ICML Workshop* (2015).

[70] KOVESI, P., ET AL. Image features from phase congruency. *Videre: Journal of computer vision research* (1999).

[71] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (2012), F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., Curran Associates, Inc.

[72] LAO, S., GONG, Y., SHI, S., YANG, S., WU, T., WANG, J., XIA, W., AND YANG, Y. Attentions help cnns see better: Attention-based hybrid image quality assessment network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).

[73] LARSON, E. C., AND CHANDLER, D. M. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging* (2010).

[74] LATHUILIÈRE, S., MESEJO, P., ALAMEDA-PINEDA, X., AND HORAUD, R. A comprehensive analysis of deep regression. *IEEE TPAMI* (2020).

[75] LECUN, Y., BOTTOU, L., BENGIO, Y., AND HAFFNER, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* (1998).

[76] LEDIG, C., THEIS, L., HUSZÁR, F., CABALLERO, J., CUNNINGHAM, A., ACOSTA, A., AITKEN, A., TEJANI, A., TOTZ, J., WANG, Z., ET AL. Photo-realistic single

image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017).

[77] LI, D., JIANG, T., LIN, W., AND JIANG, M. Which has better visual quality: The clear blue sky or a blurry animal? *IEEE Transactions on Multimedia* (2018).

[78] LI, J., MANTIUK, R., WANG, J., LING, S., AND LE CALLET, P. Hybrid-mst: A hybrid active sampling strategy for pairwise preference aggregation. *Advances in neural information processing systems* (2018).

[79] LIN, H., HOSU, V., AND SAUPE, D. Kadid-10k: A large-scale artificially distorted iqa database. In *QoMEX* (2019), IEEE.

[80] LIN, K.-Y., AND WANG, G. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In *IEEE CVPR* (2018).

[81] LIN, T.-Y., DOLLÁR, P., GIRSHICK, R., HE, K., HARIHARAN, B., AND BELONGIE, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017).

[82] LOEBICH, C., WUELLER, D., KLINGEN, B., AND JAEGER, A. Digital camera resolution measurements using sinusoidal siemens stars. In *Digital Photography III* (2007), International Society for Optics and Photonics.

[83] LUAN, F., PARIS, S., SHECHTMAN, E., AND BALA, K. Deep photo style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017).

[84] MADHUSUDANA, P. C., BIRKBECK, N., WANG, Y., ADSUMILLI, B., AND BOVIK, A. C. Image quality assessment using contrastive learning. *TIP* (2022).

[85] MAÎTRE, H. *From photon to pixel: the digital camera handbook*. John Wiley & Sons, 2017.

[86] MANTIUK, R. K., TOMASZEWSKA, A., AND MANTIUK, R. Comparison of four subjective methods for image quality assessment. In *Computer graphics forum* (2012), Wiley Online Library.

[87] MARZILIANO, P., DUFAUX, F., WINKLER, S., AND EBRAHIMI, T. Perceptual blur and ringing metrics: application to jpeg2000. *Signal processing: Image communication* (2004).

[88] MATHERON, G. *Random sets and integral geometry*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. 1975.

[89] MEYER, J. S., INGERSOLL, C. G., MCDONALD, L. L., AND BOYCE, M. S. Estimating uncertainty in population growth rates: jackknife vs. bootstrap techniques. *Ecology* (1986).

[90] MIKHAILIUK, A., WILMOT, C., PEREZ-ORTIZ, M., YUE, D., AND MANTIUK, R. K. Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), IEEE.

[91] MILOVANOVIĆ, M., TARTAGLIONE, E., CAGNAZZO, M., AND HENRY, F. Learn how to prune pixels for multi-view neural image-based synthesis. *arXiv preprint arXiv:2305.03572* (2023).

[92] MITTAL, A., MOORTHY, A. K., AND BOVIK, A. C. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* (2012).

[93] MITTAL, A., SOUNDARARAJAN, R., AND BOVIK, A. C. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters* (2012).

[94] MOORTHY, A. K., AND BOVIK, A. C. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing* (2011).

[95] PEREZ-ORTIZ, M., AND MANTIUK, R. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint* (2017).

[96] PEREZ-ORTIZ, M., MIKHAILIUK, A., ZERMAN, E., HULUSIC, V., VALENZISE, G., AND MANTIUK, R. K. From pairwise comparisons and rating to a unified quality scale. *ICIP* (2019).

[97] PHILLIPS, J. B., AND ELIASSON, H. *Camera image quality benchmarking*. John Wiley & Sons, 2018.

[98] PONOMARENKO, N., JIN, L., IEREMEIEV, O., LUKIN, V., EGIAZARIAN, K., ASTOLA, J., VOZEL, B., CHEHDI, K., CARLI, M., BATTISTI, F., ET AL. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication* (2015).

[99] PONOMARENKO, N., LUKIN, V., ZELENSKY, A., EGIAZARIAN, K., CARLI, M., AND BATTISTI, F. Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics* (2009).

[100] POPEL, M., TOMKOVA, M., TOMEK, J., KAISER, Ł., USZKOREIT, J., BOJAR, O., AND ŽABOKRTSKỲ, Z. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications* (2020).

[101] PRASHNANI, E., CAI, H., MOSTOFI, Y., AND SEN, P. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018).

[102] PUMAROLA, A., CORONA, E., PONS-MOLL, G., AND MORENO-NOGUER, F. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 10318–10327.

[103] RADFORD, A., KIM, J. W., HALLACY, C., RAMESH, A., GOH, G., AGARWAL, S., SASTRY, G., ASKELL, A., MISHKIN, P., CLARK, J., ET AL. Learning transferable visual models from natural language supervision. In *International conference on machine learning* (2021), PMLR.

[104] REISENHOFER, R., BOSSE, S., KUTYNIOK, G., AND WIEGAND, T. A haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication* (2018).

[105] RENAUDIN, M., VLACHOMITROU, A.-C., FACCIOLO, G., HAUSER, W., SOMMELET, C., VIARD, C., AND GUICHARD, F. Towards a quantitative evaluation of multi-imaging systems. In *IQSP* (2017).

[106] ROMBACH, R., BLATTMANN, A., LORENZ, D., ESSER, P., AND OMMER, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).

[107] RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A., BERNSTEIN, M., BERG, A. C., AND FEI-FEI, L. ImageNet Large Scale Visual Recognition Challenge. *IJCV* (2015).

[108] SAAD, M. A., BOVIK, A. C., AND CHARRIER, C. A dct statistics-based blind image quality index. *IEEE Signal Processing Letters* (2010).

[109] SAJJADI, M. S., SCHOLKOPF, B., AND HIRSCH, M. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE international conference on computer vision* (2017).

[110] SATTLER, T., LEIBE, B., AND KOBBELT, L. Scramsac: Improving ransac's efficiency with a spatial consistency filter. In *ICCV* (2009), IEEE.

[111] SCHÖLKOPF, B., PLATT, J., AND HOFMANN, T. Trueskill™: A bayesian skill rating system.

[112] SEO, H. J., AND MILANFAR, P. Static and space-time visual saliency detection by self-resemblance. *Journal of vision* (2009).

[113] SHEIKH, H. R., BOVIK, A. C., AND CORMACK, L. No-reference quality assessment using natural scene statistics: Jpeg2000. *IEEE Transactions on image processing* (2005).

[114] SHEIKH, H. R., SABIR, M. F., AND BOVIK, A. C. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE TIP* (2006).

[115] SHEIKH, H. R., WANG, Z., CORMACK, L., AND BOVIK, A. C. Blind quality assessment for jpeg2000 compressed images. In *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002.* (2002), IEEE.

[116] SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[117] SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556* (2014).

[118] SU, S., YAN, Q., ZHU, Y., ZHANG, C., GE, X., SUN, J., AND ZHANG, Y. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE CVPR* (2020).

[119] SUMNER, R. C., BURADA, R., AND KRAM, N. The effects of misregistration on the dead leaves crosscorrelation texture blur analysis. *Electronic Imaging* (2017).

[120] TALEBI, H., AND MILANFAR, P. Nima: Neural image assessment. *IEEE transactions on image processing* (2018).

[121] THOMEE, B., SHAMMA, D. A., FRIEDLAND, G., ELIZALDE, B., NI, K., POLAND, D., BORTH, D., AND LI, L.-J. Yfcc100m: The new data in multimedia research. *Communications of the ACM* (2016).

[122] TIMOFTE, R., DE SMET, V., AND VAN GOOL, L. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE international conference on computer vision* (2013), pp. 1920–1927.

[123] TONG, H., LI, M., ZHANG, H.-J., AND ZHANG, C. No-reference quality assessment for jpeg2000 compressed images. In *2004 International Conference on Image Processing, 2004. ICIP'04.* (2004), IEEE.

[124] TONG, Y., KONIK, H., CHEIKH, F., AND TREMEAU, A. Full reference image quality assessment based on saliency map analysis. *Journal of Imaging Science and Technology* (2010).

[125] TUKEY, J. W. Comparing individual means in the analysis of variance. *Biometrics* (1949).

[126] TWORSKI, M., AND LATHUILIÈRE, S. Test your samples jointly: Pseudo-reference for image quality evaluation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023), IEEE, pp. 1–5.

[127] TWORSKI, M., LATHUILIÈRE, S., BELKARFA, S., FIANDROTTI, A., AND CAGNAZZO, M. Dr2s : Deep regression with region selection for camera quality evaluation. In *ICPR* (2020).

[128] TWORSKI, M., POCHON, B., AND LATHUILIÈRE, S. Camera quality assessment in real-world conditions. *Available at SSRN 4166549* (2022).

[129] VAN ZWANENBERG, O., TRIANTAPHILLIDOU, S., JENKIN, R., AND PSARROU, A. Edge detection techniques for quantifying spatial imaging system performance and image quality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2019).

[130] VARGA, D. A combined full-reference image quality assessment method based on convolutional activation maps. *Algorithms* (2020).

[131] VARGA, D. Composition-preserving deep approach to full-reference image quality assessment. *Signal, Image and Video Processing* (2020).

[132] WANG, Q., WANG, Z., GENOVA, K., SRINIVASAN, P. P., ZHOU, H., BARRON, J. T., MARTIN-BRUALLA, R., SNAVELY, N., AND FUNKHOUSER, T. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 4690–4699.

[133] WANG, X., YU, K., WU, S., GU, J., LIU, Y., DONG, C., QIAO, Y., AND CHANGE LOY, C. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops* (2018).

[134] WANG, Z., BOVIK, A. C., SHEIKH, H. R., AND SIMONCELLI, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* (2004).

[135] WANG, Z., SHEIKH, H. R., AND BOVIK, A. C. No-reference perceptual quality assessment of jpeg compressed images. In *Proceedings. International conference on image processing* (2002), IEEE.

[136] WANG, Z., SIMONCELLI, E. P., AND BOVIK, A. C. Multiscale structural similarity for image quality assessment.

[137] WUELLER, D., MATSUI, A., AND KATOH, N. Visual noise revision for ISO 15739. *Electronic Imaging* (Jan. 2019).

[138] XU, J., ZHANG, L., AND ZHANG, D. A trilateral weighted sparse coding scheme for real-world image denoising. In *Proceedings of the European conference on computer vision (ECCV)* (2018).

[139] YANG, J., WRIGHT, J., HUANG, T. S., AND MA, Y. Image super-resolution via sparse representation. *IEEE transactions on image processing* (2010).

[140] YE, P., KUMAR, J., KANG, L., AND DOERMANN, D. Unsupervised feature learning framework for no-reference image quality assessment. In *IEEE CVPR* (2012), IEEE.

[141] YE, P., KUMAR, J., KANG, L., AND DOERMANN, D. Real-time no-reference image quality assessment based on filter learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013).

[142] YING, Z., NIU, H., GUPTA, P., MAHAJAN, D., GHADIYARAM, D., AND BOVIK, A. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020).

[143] YU, N., SHEN, X., LIN, Z., MECH, R., AND BARNES, C. Learning to detect multiple photographic defects. In *IEEE WACV* (2018), IEEE.

[144] ZENG, H., ZHANG, L., AND BOVIK, A. C. Blind image quality assessment with a probabilistic quality representation. In *IEEE International Conference on Image Processing (ICIP)* (2018), IEEE.

[145] ZERMAN, E., HULUSIC, V., VALENZISE, G., MANTIUK, R., AND DUFAUX, F. The relation between mos and pairwise comparisons and the importance of cross-content comparisons. In *Human Vision and Electronic Imaging Conference, IS&T International Symposium on Electronic Imaging (EI 2018)* (2018).

[146] ZHANG, L., GU, Z., AND LI, H. Sdsp: A novel saliency detection method by combining simple priors. In *2013 IEEE international conference on image processing* (2013), IEEE.

[147] ZHANG, L., SHEN, Y., AND LI, H. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing* (2014).

[148] ZHANG, L., ZHANG, L., AND BOVIK, A. C. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing* (2015).

[149] ZHANG, L., ZHANG, L., MOU, X., AND ZHANG, D. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing* (2011).

[150] ZHANG, R., ISOLA, P., EFROS, A. A., SHECHTMAN, E., AND WANG, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018).

[151] ZHANG, R., ISOLA, P., EFROS, A. A., SHECHTMAN, E., AND WANG, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018).

[152] ZHANG, W., LIU, Y., DONG, C., AND QIAO, Y. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019).

[153] ZHANG, W., MA, K., YAN, J., DENG, D., AND WANG, Z. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology* (2018).

[154] ZHANG, X., WANDELL, B. A., ET AL. A spatial extension of cielab for digital color image reproduction. In *SID international symposium digest of technical papers* (1996), Citeseer.

[155] ZHENG, H., YANG, H., FU, J., ZHA, Z.-J., AND LUO, J. Learning conditional knowledge distillation for degraded-reference image quality assessment. In *IEEE ICCV* (2021).

[156] ZHOU, B., KHOSLA, A., LAPEDRIZA, A., OLIVA, A., AND TORRALBA, A. Learning deep features for discriminative localization. In *CVPR* (2016).

[157] ZHU, H., LI, L., WU, J., DONG, W., AND SHI, G. Metaiqa: Deep meta-learning for no-reference image quality assessment. In *CVPR* (2020).

ECOLE DOCTORALE

Titre : Analyse automatique de critères de qualité d'image sur scènes naturelles par réseaux de neurones profonds

Mots clés : Evaluation de la qualité d'image, Texture, Apprentissage profond, Adaptation de domaine

Résumé : Alors que les caméras de smartphones sont devenues plus répandues que les systèmes de caméras traditionnels, la demande de mesures précises a augmenté. Cette thèse de doctorat propose d'utiliser des systèmes d'apprentissage profond pour évaluer les critères de qualité d'image spécifiques à l'évaluation des caméras de smartphones, et plus particulièrement l'évaluation de la texture. Cette thèse aborde plusieurs limites des méthodes actuelles d'évaluation de la qualité d'image pour les caméras de smartphones. Les systèmes d'apprentissage profond se heurtent à la complexité de calcul due à la haute résolution des images de smartphones, et la réduction de la taille entraînerait une perte d'informations pour l'évaluation de la préservation du bruit et des détails. Par conséquent, il est essentiel de trouver les régions d'image pertinentes pour évaluer les attributs d'un appareil photo afin d'atténuer ces problèmes. En outre, le manque d'ensembles de données appropriés entrave le développement de méthodes basées sur l'apprentissage visant à évaluer les appareils photo de smartphones. En outre, lors de la comparaison des appareils photo, il est essentiel de capturer le même contenu pour faciliter la comparaison directe. Dans les protocoles standard d'évaluation comparative des appareils photo, plusieurs photos sont prises à partir du même contenu. Ce cadre s'écarte des approches traditionnelles d'apprentissage automatique, dans lesquelles les données d'apprentissage et de test sont supposées être indépendantes et identiquement distribuées (*iid*). Cependant, la nature non indépendante de nos données est souvent négligée dans la littérature sur l'évaluation de la qualité des images.
Pour relever ces défis, cette recherche apporte plusieurs contributions : **(i)**. Une méthode de sélection des régions est introduite pour détecter automatiquement les régions pertinentes pour l'évaluation d'attributs de qualité spécifiques. En adaptant la méthode de la carte d'activation de classe à un problème de régression, nous surpassons les approches traditionnelles basées sur des mires spécialisées pour évaluer la qualité de la texture et permettre l'utilisation de méthodes d'apprentissage profond sur des graphiques pris en laboratoire. Dans ce travail, nous utilisons la qualité de la texture comme exemple illustratif des attributs de qualité de la caméra. Cependant, notre méthodologie est conçue pour s'appliquer également à d'autres attributs, tels que le bruit. **(ii)** Un nouvel ensemble de données in-the-wild est créé pour refléter avec précision le mélange complexe de défauts que l'on trouve couramment dans les images d'appareils photo de smartphones et pour refléter le scénario de l'évaluation comparative des appareils photo, dans lequel plusieurs scènes différentes sont filmées par plusieurs appareils photo. Cet ensemble de données, annoté par des comparaisons par paire, nous permet d'effectuer une large évaluation de différentes méthodes dans différents scénarios pratiques, établissant des lignes directrices pour l'utilisation de systèmes d'apprentissage profond pour l'évaluation de la qualité des appareils photo. **(iii)** Nous introduisons une nouvelle configuration et une nouvelle méthode d'évaluation de la qualité des images qui vont au-delà de l'hypothèse *iid* traditionnelle. Nous considérons plusieurs images de qualité variable du même contenu disponibles au moment du test. Nous utilisons la spécificité de ce cadre d'estimation de la qualité de la caméra pour améliorer la précision de la prédiction de la qualité en introduisant une pseudo-référence basée sur le batch qui nous permet d'utiliser des méthodes de référence complète dans le cadre sans référence.

Title : Automatic analysis of image quality criteria in natural scenes using deep neural networks

Keywords : Image quality assessment, Texture, Deep learning, Domain adaptation

Abstract : As smartphone cameras became more prevalent than traditional camera systems, the demand for precise measurements increased. This Ph.D. dissertation proposes using deep learning systems to evaluate image-quality criteria specific to smartphone camera evaluation, and more specifically texture evaluation. This dissertation addresses several limitations in current image-quality assessment methods for smartphone cameras. Deep learning systems struggle with computational complexity due to high-resolution smartphone images, and downsizing would lead to information loss for evaluating noise and details preservation. Consequently, it is essential to find the relevant image regions to assess a camera attribute to alleviate these problems. Additionally, the lack of suitable datasets hinders the development of learning-based methods aimed at benchmarking smartphone cameras. Furthermore, when comparing cameras, it is essential to capture the same content to facilitate direct comparison. In standard camera benchmarking protocols, multiple shots are collected from the same content. This setting deviates from traditional machine learning approaches, where training and test data are assumed to be independent and identically distributed (*iid*). However, the non-independent nature of our data is frequently overlooked in the image-quality assessment literature.
To overcome these challenges, this research introduces several contributions: **(i)** A region selection method is introduced to automatically detect relevant regions for evaluating specific quality attributes. Adapting the class activation map method for a regression problem, we outperform traditional chart-based approaches in evaluating texture quality and permitting the usage of deep learning methods on charts shot in laboratory conditions. In this work, we use texture quality as an illustrative example of camera quality attributes. However, our methodology is designed to be applicable to other attributes, such as noise, as well. **(ii)** A new in-the-wild dataset is created to accurately reflect the complex mixture of defects commonly found in smartphone camera images and reflect the scenario of camera benchmarking, where several different scenes are shot by multiple camera devices. This dataset, annotated through pairwise comparisons, allows us to perform a large evaluation of different methods in different practical scenarios, setting guidelines for the usage of deep learning systems for camera quality evaluation. **(iii)** We introduce a new image quality assessment setup and method where we go beyond the traditional *iid* assumption. We consider multiple images with varying quality of the same content available at test time. We use the specificity of this camera quality estimation setting to enhance the quality prediction accuracy by introducing a batch-based pseudo-reference which allows us to use full-reference methods in the no-reference setting.

Institut Polytechnique de Paris
91120 Palaiseau, France