HAL open science

Selected Contributions in Video Compression, Quality of Experience and Analysis

Giuseppe Valenzise

▶ To cite this version:

Giuseppe Valenzise. Selected Contributions in Video Compression, Quality of Experience and Analysis. Signal and Image processing. université paris sud, 2019. tel-02996772

HAL Id: tel-02996772
https://hal.science/tel-02996772v1

Submitted on 9 Nov 2020

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

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

Université Paris-Saclay

Habilitation à diriger des recherches

Selected Contributions in Video Compression, Quality of Experience and Analysis

Giuseppe Valenzise

18 of December, 2019

Preface

This manuscript summarizes some of the research activities I conducted after obtaining my Phd degree from Politecnico di Milano in February 2011. Soon after my graduation, I moved to the Laboratoire Traitement et Communication de l'Information (LTCI), Telecom Paristech, initially as a post-doc researcher, and since 2012 as a permanent CNRS researcher ("chargé de recherches"). This position gave me a great freedom to study multiple aspects of visual information processing concurrently, and to take risks in exploring different topics from those I worked on in my PhD thesis. Since the end of 2016, I carry out my research at the Laboratoire des Signaux et Systèmes (L2S), CentraleSupelec, Universitè Paris Sud, Universitè Paris Saclay.

As mentioned above, my research covers several aspects of visual information processing. This is a rich and variegate field, ranging from image/video acquisition to coding and transmission, models of perception, quality of experience, interactivity, security and analysis. In this context, my research mainly focuses on three aspects. The first is finding efficient representations for **image/video compression**: this is especially relevant for new and emerging video formats, such as high dynamic range (HDR) video, light-field content, point clouds, etc. A second research theme is assessing and modeling the perception of the **quality of experience** (QoE) for image/video content, in particular for new and richer video representations. Indeed, this enables to improve existing compression algorithms and to design new ones, taking into account the perceived quality of coding artifacts. Also, measuring the QoE enables to gauge the added value of these media in pursuing a truly realistic and immersive video experience. Finally, the third research thread is about **image/video analysis**. Oftentimes, image/video content is not intended for entertainment and might be input to other, possibly automatic, tasks such as object detection, recognition or tracking. This entails a more general notion of perception, which is becoming increasingly important in many applications, from video surveillance to autonomous driving. In this respect, my focus is mainly on how to process visual information in order to facilitate further analysis tasks, again with a particular emphasis on richer content representations.

This manuscript is organized in three parts. The first part consists of an extended curriculum vitae, with a focus on the research activities I have supervised/coordinated. In the second part, I provide a selection of my past research contributions, according to the three main themes mentioned above. Each chapter and section can be read in a "multi-resolution" fashion: the introduction provides an overall outline of the contributions; some details and the most significant results are reported in the rest of the text; and finally, for a complete presentation of the work, the readers are pointed to the original papers where the contributions have been published. Finally, the third and last part of the manuscript provides some insights on my future research perspective in this domain.

Contents

Preface **i**

I Extended Curriculum Vitae **2**

II Summary of research activities **11**

1 Video compression **12**

1.1 Depth video compression using *Don't Care Regions* 12

1.1.1 Definition of Don't Care Region (DCR) 13

1.1.2 Application to temporal coding 14

1.1.3 Experimental results 15

1.2 Optimal tone mapping for HDR video compression 16

1.2.1 Content Adaptive Tone Mapping Operator 17

1.2.2 Experimental results 21

1.3 Enhancing Intra prediction using context-based learning 23

1.3.1 Network 24

1.3.2 Experimental results 25

1.4 Learning-based coding of point cloud geometry 26

1.4.1 Proposed PCC coding scheme 28

1.4.2 Experimental results 29

2 Quality of Experience **31**

2.1 An extensive evaluation of HDR fidelity metrics 32

2.1.1 Datasets and quality metrics 33

2.1.2 Statistical analysis 36

2.1.3 Discriminability analysis 37

2.2 Blind HDR quality estimation disentangling perceptual and noisy features . . . 40

2.2.1 Proposed model 40

2.2.2 Two-step training procedure 43

2.2.3 Experimental results 43

2.3 Perceived dynamic range for HDR images 46

2.3.1 Subjective dataset 47

2.3.2 Considered image features 49

2.3.3 PDR predictor model 50

2.4 Towards a unified quality scale fusing rating and ranking measures 51

2.4.1 Observer model 52

2.4.2 Psychometric scaling 53

2.4.3 The relation between MOS and PWC and the importance of cross-content comparisons 56

2.4.4 Combining rating and pairwise comparisons 59

3 Image and video analysis **62**

3.1 Local features for RGBD image matching 63

3.1.1 Background concepts on image matching 63

3.1.2 Keypoint extraction based on a RGBD scale space 65

3.1.3 TRISK: local features extraction for RGBD content matching 71

3.2 Learning-based tone mapping for robust image matching under illumination changes 78

3.2.1 Design of OpTMO 79

3.2.2 Experimental results 82
3.3 Detection of inverse tone mapping in HDR images 83
3.3.1 Forensic analysis based on Fisher scores 85
3.3.2 Experimental results 86
3.4 Predicting subjectivity in image aesthetic assessment 87
3.4.1 Subjectivity measures 88
3.4.2 Prediction of subjectivity 90

III Future research perspectives 93

4 Towards effective representations for immersive visual communication 94
4.1 Learning good representations for video compression and quality assessment . 95
4.1.1 Deep generative models for video coding 95
4.1.2 Representation learning in quality assessment 97
4.2 New methodologies for Quality of Experience 98
4.3 Immersive visual communication 99

Appendix 100

A List of publications 101

Part I

Extended Curriculum Vitae

General information

Name:	Giuseppe
Surname:	Valenzise
Birth:	28th of December 1982, Monza, Italy
Nationality:	Italian
E-mail:	giuseppe.valenzise@l2s.centralesupelec.fr
Website:	http://webpages.l2s.centralesupelec.fr/perso/giuseppe.valenzise/

Current position

Position (since Oct. 2012):	Chargé de recherches CNRS (researcher)
Affiliation (since Nov. 2016):	UMR 8506, Laboratoire des Signaux et Systèmes (L2S) – CNRS, CentraleSupélec, Université Paris-Sud
Address:	3 rue Joliot-Curie, 91192 Gif-sur-Yvette Cedex, France

Previous experience

Oct. 2012 - Oct. 2016:	Chargé de recherches CNRS at LTCI, Telecom Paristech, Paris, France
Jul. 2011 - Sep. 2012:	Post-doc researcher at Telecom Paristech, Paris, France
Feb. 2011 - Jun. 2011:	Post-doc researcher at Politecnico di Milano, Italy
Jan. 2009 - Aug. 2009:	Visiting researcher at University of Southern California, Los Angeles, USA

Education

Feb. 2011:	PhD in Information Technology, Politecnico di Milano, Italy. *Thesis: "No-reference and reduced-reference methods for multimedia forensics and quality assessment"*
Apr. 2007:	MSc in Computer Engineering, Politecnico di Milano, Italy

Supervision activities

Former PhD theses

1. **Paul Lauga**, "High Dynamic Range (HDR) representations for Digital Video" (with F. Dufaux, starting from June 2012; thesis defended on December 3, 2015)

   HDR processing allows to capture and represent a scene with a greater level of realism than conventional, low dynamic range (LDR) images. This representation brings new challenges such as encoding and evaluating the quality of HDR content, or converting

HDR to LDR and vice versa. In his thesis, Dr. Lauga proposed several new compression methods adapted to HDR images and video sequences. A first method segments the HDR image into dark and bright regions, in order to preserve more details of the original picture [94, 95]. The next two methods propose to include two regularization terms for the spatial [96] and temporal [137, 138] complexity, respectively, in order to optimize the rate-distortion trade-off. These compression methods have been compared with those in the state of the art using several HDR quality metrics. The effectiveness of these metrics in the case of HDR image compression has been validated through a subjective study [189]. A final study evaluates subjectively the performance of the application of different expansion algorithms on LDR video sequences in order to display them on a HDR display [26].

Dr. Lauga is currently employed as a post-doc researcher at Hypervision technology, Paris, in collaboration with Centrale Marseille. Previously, he worked as a post-doc researcher at Institut Fresnel, Marseille.

2. **Maxim Karpushin**, "Local features for RGBD image matching under viewpoint changes" (with F. Dufaux, starting October 2013; thesis defended on November 3, 2016)

   Texture+depth (RGBD) visual content acquisition offers new possibilities for different classical problems in vision, robotics and multimedia. In his thesis, Dr. Karpushin has addressed the task of establishing local visual correspondences in images, which is a basic operation underlying numerous higher-level problems, including object tracking, visual odometry, multimedia indexing and visual search. The local correspondences are commonly found through local visual features. While these have been exhaustively studied for traditional images, little has been done for the case of RGBD content. Yet, depth can provide useful information to increase the robustness of local features to one of the major difficulties in image matching, i.e., drastic viewpoint changes. Based on this observation, Dr. Karpushin contributed several new approaches of keypoint detection and descriptor extraction, that preserve the conventional degree of keypoint covariance and descriptor invariance to in-plane visual deformations, but aim at improved stability to out-of-plane (3D) transformations in comparison to existing texture-only and texture+depth local features. More specifically, Dr. Karpushin initially proposed to normalize the local descriptor sampling patterns in order to adapt to the local geometry of objects in the scene, by either using planar approximations [70] or by directly sampling the surface manifold [71]. The latter approach may provide better invariance to descriptor extraction, but entails a higher computational cost. Therefore, Dr. Karpushin focused on the former strategy, formalizing the concept of *local adaptive axes* [76, 77]. On the other hand, Dr. Karpushin revisited a classical construction of local keypoint detection: scale spaces. Based on the axiomatic definition of scale space, Dr. Karpushin derived a stable numerical diffusion process and proved that this engenders a scale space, satisfying the scale space axioms [72, 74]. He also proposed a simpler and faster approximation to the scale space construction [75]. Finally, he developed a fast and accurate quadratic approximation to Gaussian smoothing, required for generating scale spaces in 2D, based on integral images [73]. Dr. Karpushing evaluated these proposed methods using application-level scenarios on RGBD datasets acquired with Kinect sensors.

Dr. Karpushin is currently a senior R&D engineer at GoPro, France.

3. **Emin Zerman**, "Assessment and analysis of high dynamic range video quality" (with F. Dufaux, starting February 2015; thesis defended on January 19, 2018)

   High dynamic range video quality assessment presents a number of differences and challenges with respect to conventional video quality assessment, e.g., the impact of brighter displays on the perception of distortion and colors, and how to take into account these phenomena when computing video quality metrics. In his PhD thesis, Dr. Zerman identified and targeted some of these challenges. Initially, he considered the effects of display rendering on HDR image quality assessment. To this end, he proposed a new display shading algorithm [209], and he carried out a subjective study to investigate the effect of different shading methods on both subjective scores and objective metrics [208]. Afterwards, Dr. Zerman studied the effect of using different color spaces on HDR video compression performance [206]. Interestingly, these studies found that quality (fidelity) scores for compression artifacts are quite robust to both different rendering schemes and color transformations, suggesting strong similarities to low dynamic range (LDR) quality evaluation. In order to further study these similarities, and specifically how existing LDR quality metrics can be adapted to predict HDR quality, Dr. Zerman conducted an extensive validation of existing fidelity measures [210], including metrics specifically designed for HDR as well as metrics originally proposed for LDR and adapted to HDR by means of a proper perceptual encoding. This study collects aligned opinion scores from several previous quality surveys, and proposes a new one, yielding the largest available benchmark of HDR image fidelity metrics.

   In the last part of his PhD thesis, Dr. Zerman focused on more methodological aspects of gauging subjective quality scores. More specifically, in collaboration with Dr. Rafal Mantiuk (University of Cambridge, UK), he studied the relationship between rating scores (such as MOS, mean opinion scores), and pairwise comparison scores [207]. This work is currently being extended towards defining a unified quality scale using scores deriving from different experiments.

   Dr. Zerman is currently working as a post-doctoral researcher at the V-SENSE lab, Trinity College of Dublin, Ireland.

4. **Aakanksha Rana**, "High dynamic range image analysis" (with F. Dufaux, starting April 2015; thesis defended on March 15, 2018)

   Drastic illumination changes are one of the greatest challenges for image matching and, in general, image analysis. One way to face this challenge, which was the subject of Dr. Rana's Phd thesis, consists in employing a richer visual representation, such as the one provided by HDR imaging, in order to capture more details both in the bright and dark areas of the scene. In principle, HDR represents the scene radiance, and should provide invariance to illumination changes. In practice, the HDR content needs to be properly processed in order to obtain this invariance. Traditionally, *tone mapping* operators (TMO) have been proposed to separate scene illumination from object reflectance. However, these algorithms have been conventionally designed and tuned to maximize the visual experience of human observers, rather than for facilitating machine tasks. Dr. Rana

began her thesis by exploring which kind of HDR representation (linear radiance values, different tone mapping methods or fixed transfer function to compress the luminance range) best fits the task of image matching [151, 152]. She found that dynamic range compression is necessary to obtain more stable and descriptive features, while linear light is not adapted to current feature extraction pipelines. However, no single existing HDR representation gives optimal performance in general, which suggests potential gains in optimizing a tone mapping operator for a given content and for the specific task of image matching [153]. Based on these observations, Dr. Rana proposed a learning approach to derive the optimal TMO for a given scene, by collecting a dataset of pictures with different illumination conditions. Afterwards, she used this dataset to learn a local parametric bilateral filter able to optimize both keypoint detection [154] and matching [155, 156]. In the last part of her thesis, Dr. Rana proposed an end-to-end optimized tone mapping operator, based on a conditional generative adversarial network (cGAN). This work is under submission at the time of this writing.

Dr. Rana is currently working as a post-doctoral researcher at Harward Medical School, Boston, USA. Previously, she worked as a post-doctoral researcher at the V-SENSE lab, Trinity College of Dublin, Ireland.

Former post-doctoral fellows

1. **Cagri Ozcinar**, 1 year

   Dr. Ozcinar worked in collaboration with Paul Lauga to the definition and implementation of a temporally coherent tone mapping operator for scalable HDR video compression [137, 138]. The proposed TMO consists in formulating a convex optimization problem to minimize HDR video reconstruction distortion, while at the same time taking into account the temporal smoothness of the luminance of motion trajectories.

   Dr. Ozcinar is currently working as a post-doctoral researcher at the V-SENSE lab, Trinity College of Dublin, Ireland.

2. **Wei Fan**, 1 year

   Dr. Fan post-doctoral research focused on the forensic analysis of high dynamic range content. In particular, one important difference between HDR and LDR pictures is that the former might be obtained by fusing multiple pictures with different exposure (mHDR) or by artificially boosting the dynamic range of an LDR picture, through an operation called *inverse tone mapping* (iHDR). Dr. Fan proposed a forensic feature to distinguish mHDR from iHDR pictures. The feature, based on the Fisher score of a Gaussian mixture model of local pixel dependencies, was showed to be discriminative and robust to a number of different inverse tone mapping methods [34, 35].

   After her post-doc experience in Telecom Paristech, Dr. Fan has worked as post-doc researcher at Dartmouth University, US, and she is currently at Google, Mountain View, as a senior engineer.

3. **Vedad Hulusic**, 2 years and 3 months

   Dr. Hulusic mainly worked on measuring and modeling the perception of dynamic range in HDR pictures. He built a (publicly available) subjectively annotated dataset of HDR pictures with dynamic range judged by people on a rating scale [52]. Afterwards, he used this dataset to develop a computational model to predict the perceived dynamic range taking into account some geometric aspects of light distribution in the picture, linking it to previously studied concepts in lightness perception [48, 49]. He further extended this study to consider the case of stimuli with no semantic information [50]. In the context of the 4EVER2 project, Dr. Hulusic conducted a study on the combined effect of Ultra High Definition (UHD) and High Frame Rate (HFR) video formats to the quality of experience [51]. He also co-supervised an intern student to carry out a subjective study on the video quality for tile-based spatial adaptation in HTTP adaptive streaming.

   Dr. Hulusic is currently a senior Lecturer in Games Design and Development at the Department of Creative Technology, Faculty of Science and Technology, University of Bournemouth, UK.

Current PhD students

1. **Chen Kang**, "Aesthetic video quality assessment using deep neural networks", started in October 2017; expected graduation September 2020. I am the main director of this thesis, which is co-supervised by F. Dufaux (50%)

   Predicting aesthetic judgments of video is a highly subjective process, and is far more challenging than traditional quality assessment. In her PhD thesis, Chen Kang studies how to predict aesthetic quality using deep convolutional neural networks. Deep learning approaches have been successfully used in the last few years to predict aesthetic quality in a supervised fashion, thanks to the availability of large-scale subjectively annotated public datasets such as AVA. In the first part of the thesis, Chen Kang focused on quantifying and estimating the *subjectivity* of aesthetic scores [67]. In fact, images having similar average ratings might display very different degrees of consensus among human observers. In this context, she proposed several measures to quantify this consensus, showing that it is more efficient to directly predict these measures rather than predicting the distribution of the scores, which seems to be an intrinsically more difficult task. Existing aesthetic datasets are very noisy, and subjective scores do not reflect only the aesthetic value of pictures but are influenced by many external or personal factors (photographic challenge, topic, interestingness, popularity, etc.). Therefore, in the second part of the thesis we focus on building a new aesthetic datasets, which will be collected in a more disciplined way, e.g., by clearly defining aesthetic attributes and training raters to their intepretation. This will enable to improve aesthetic prediction and interpretability, and to facilitate tasks such as automatic image enhancement.

2. **Maurice Quach**, "Compression and Quality Assessment of Point Cloud", started in October 2018; expected graduation September 2021. Co-supervised with F. Dufaux at 50%.

   The PhD thesis of Maurice Quach focuses on the efficient compression of point cloud

video, and on the estimation of the perceptual quality of compressed point clouds. In particular, in the first part of his thesis he has focused on the lossy compression of point cloud geometry, proposing a learning-based framework based on auto-encoders [150].

3. **Milan Stepanov**, "View extraction and coding for light field imaging", started in December 2018; expected graduation November 2021. Co-supervised with F. Dufaux at 50%.

   The goal of Milan Stepanov's thesis is to investigate new coding approaches for light field video. In particular, he will explore learning-based solutions and adapt them to the specific structure of light field data.

Current Post-doctoral fellows

1. **Li Wang**, started February 2018

   Dr. Wang conducts her post-doctoral research on video coding optimization using deep generative models. In particular, she studies how deep generative models can be employed to optimize typical coding tools, such as spatial/temporal prediction; and entropy coding, which account for a large part of the efficiency of current video coding techniques. Specifically, Li Wang has proposed an encoder-decoder convolutional network able to reduce the energy of the residuals of HEVC intra prediction, by leveraging the available context of previously decoded neighboring blocks [194]. Currently, she is studying how to extend this scheme to temporal prediction, and how to optimize the architecture by embedding a non-supervised classification of coding block prior to prediction in the system.

Teaching

- Deep Learning for Multimedia – Master Multimedia and Networking, Paris Saclay (co-organizer of the class, 3 hours of teaching). 2018
- Image Processing – Master 3IR, University Paris 13 (20 hours). 2017
- SI222: Techniques de Compression. Telecom Paris, Master (3 hours). 2012-2017
- SI350: Vidéo numérique et multimédia. Telecom Paris, Master (3 hours). 2012-2017
- Télévision numérique: distribution, services et systèmes. Telecom Paris, Formation continue (6 hours). 2012-2016
- Projet d'application final (PAF), Telecom Paris (approx. 20 hours of student mentoring). 2014-2016

Scientific production and dissemination

My complete scientific production is reported in the appendix and includes:

- 18 published journal papers;

- 60 published conference papers;
- 4 published book chapters.

According to Google Scholar, my h-index is equal to 17, with more than 1300 citations, on August 21, 2019.

Together with E. Reinhard (Technicolor) and F. Dufaux (CNRS), I gave two tutorials on high dynamic range video at EUSIPCO 2016 and ICIP 2016, respectively.

Research projects

- RealVision (ITN Marie-Curie), January 2018, ongoing
- ReVeRY – RichEr VidEo for Richer creativitY (ANR), December 2017, ongoing
  - Leader of the workpackage on video quality evaluation
- V-CODE – Video Coding Optimization using Deep Generative Models (Labex Digicosme 2018)
- Projet STIC Paris-Saclay, 2017 (financement post-doc)
- 4EVER2 – For Enhanced Video ExpeRience (FUI 19), 2015-2017
- PLEIN-PHARE – Projet d'améLioration bas-coût d'unE chaIne de vidéosurveillaNce Par l'exploitation de technologies HDR d'Analyse et de Restitution (FUI 18), 2015-2017
  - Leader of the workpackage on video coding and analysis
- NEVEx – The NExt Video Experience (FUI 11), 2011-2014
- PHC BOSPHORE 2016 – Backward compatible and native encoding of High Dynamic Range video and its perceptual evaluation
  - Leader of the French team
- NORAH – NO-Reference quality Assessment for High dynamic range content (PEPS JCJC INS2I 2017)
- CLUE-HDR – CoLor appearance based qUality Estimation for HDR content (PEPS JCJC INS2I 2016)

Editorial activities and participation to technical committees

I serve as associate editor for two international journals: *IEEE Transactions on Circuits and Systems for Video Technology* (since 2016) and *Elsevier Signal processing: Image communication* (since 2015). I have participated to the technical program committee of several conferences, including ICASSP, ICIP, MMSP, ICME, VCIP, EUSIPCO, etc.

I am an elected member for the term 2018-2020 of the IEEE technical committee on Multimedia Signal Processing (MMSP) and the IEEE technical committee on Image, Video and Multidimensional Signal Processing (IVMSP), as well as a member of the Special Area Team on Visual Information Processing (SAT-VIP) of EURASIP.

Awards

I have received the EURASIP early career award in 2018 for significant contributions to video coding and analysis. I was awarded the "Prime d'Excellence Scientifique" (prize of scientific excellence) by CNRS in 2017. My PhD student, Dr. Karpushin, received the 1st Prize of the Telecom Foundation for the best PhD thesis in 2017. I also received the second prize for best student paper at ICIP 2011.

Part II

Summary of research activities

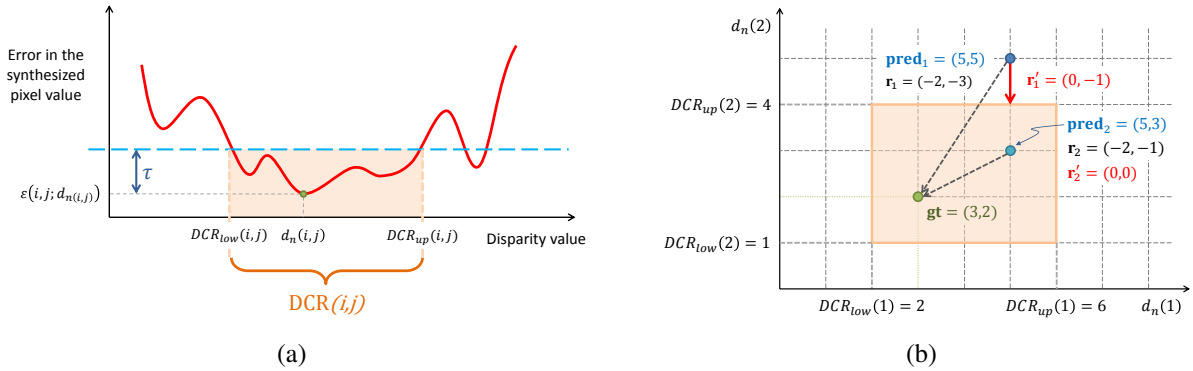
Chapter 1

Video compression

In this chapter I review my past research activity on optimization of video coding. I have been especially focusing on video representations which enable a higher degree of realism and interaction with the scene, such as high dynamic range imaging. I will start by describing in Section 1.1 a contribution on the motion prediction in depth video in the Multiview-plus-Depth (MVD) video format [188], where we exploit the fact that depth is never directly displayed to observers, but rather used to synthesize new view points. The second contribution (Section 1.2) is about the design of an optimal tone mapping operator (TMO) for high dynamic range (HDR) video coding [138]. This TMO does not only optimize the end-to-end reconstruction error of the HDR signal, but it also takes into account both spatial and temporal dependencies in order to improve the overall rate-distortion performance. Finally, more recently I have started applying machine learning tools to video compression, and to explore new video representations for virtual and mixed reality. I will conclude the chapter with some ongoing research on optimizing video compression using machine learning [194] (Section 1.3), and on the compression of point cloud geometry [150] (Section 1.4).

1.1 Depth video compression using *Don't Care Regions*

In order to enhance visual experience beyond conventional single-camera-captured video, free-viewpoint television [91] aims at enabling a richer interaction between the observer and the scene, by enabling to freely change the viewpoint as video is played back in time. A way to achieve an FTV experience consists in using elaborate arrays of closely spaced cameras (e.g., 100 cameras were used in one setup in [39]) to capture a scene of interest from multiple viewing angles. If, in addition to texture maps (RGB images), depth maps (per-pixel physical distance between scene objects and the capturing camera) are also acquired, then the observer can synthesize successive intermediate views between two camera-captured views via depth-image-based rendering (DIBR) [114] for smooth view transition. Transmitting both texture and depth maps of multiple viewpoints—a format known as *texture-plus-depth*—from server to client entails a large bit overhead, however. In this section, we address the problem of temporal coding of depth maps in texture-plus-depth format for multiview video, by introducing and employing the concept of *Don't care region*.



(a) (b)

Figure 1.1.1: *Don't Care Region.* (a) Definition of DCR for a given threshold τ. (b) Coding the residuals using DCR with a toy example with just two pixels ($d_n(1)$ and $d_n(2)$). In conventional coding, given predictor (**pred**), one aims to reconstruct the original ground truth (**gt**). However, considering DCR, it is sufficient to encode a generally smaller residual, i.e. one that enables to reconstruct a value inside or on the border of the DCR (shaded area in the picture)

This work has been conducted while I was a postdoc researcher at Telecom Paristech, and is described in greater details in the paper [188].

1.1.1 Definition of Don't Care Region (DCR)

The key observation in this work is that depth maps are not themselves directly viewed, but are only used to provide geometric information of the captured scene for view synthesis at decoder. Thus, as long as the resulting geometric error does not lead to unacceptable synthesized view quality, each depth pixel only needs to be reconstructed coarsely at decoder, e.g., within a defined tolerable range. We formalize the notion of this tolerable range per depth pixel as *don't care region* (DCR) using a threshold τ, by studying the synthesized view distortion sensitivity to the pixel value. Specifically, if a depth pixel's reconstructed value is inside its defined DCR, then the resulting geometric error will lead to distortion in a targeted synthesized view by no more than τ. Clearly a sensitive depth pixel (e.g., an object boundary pixel whose geometric error will lead to confusion between background and foreground) will have a narrow DCR, and vice versa.

More formally, a pixel $v_n(i,j)$ in texture map $\mathbf{v}_n$, with associated disparity value $d_n(i,j)$, can be obtained by a corresponding pixel in view $n+1$ through a view synthesis function $s(i,j;d_n(i,j))$. In the simplest case where the views are captured by purely horizontally shifted cameras, $s(i,j;d_n(i,j))$ corresponds to displacing a pixel in texture map $\mathbf{v}_{n+1}$ of view $n+1$ in the x-direction by an amount proportional to $d_n(i,j)$; i.e.,

$$s(i,j;d_n(i,j)) = v_{n+1}(i,j-\gamma \cdot d_n(i,j)) \tag{1.1.1}$$

where γ is a scaling factor depending on the camera spacing.

We define the *view synthesis error*, $\varepsilon(i,j;d)$, as the absolute error between the mapped-to pixel $s(i,j;d)$ in the synthesized view n and the original pixel value $v_n(i,j)$, given disparity value d

for pixel (i,j) in v_n; i.e.,

$$\varepsilon(i,j;d_n(i,j)) = |s(i,j;d_n(i,j)) - v_n(i,j)|\,. \tag{1.1.2}$$

Notice that, in general, $\varepsilon(i,j;d_n(i,j)) > 0$ because d can be quantized or noisy, and the synthesis model does not take into account disocclusions or illumination changes.

We define the *Don't Care Region* $\mathrm{DCR}(i,j) = [\mathrm{DCR}_{low}(i,j), \mathrm{DCR}_{up}(i,j)]$ as the *largest* contiguous interval of disparity values containing the ground-truth disparity $d_n(i,j)$, such that the view synthesis error for any point of the interval is smaller than $\varepsilon(i,j;d_n(i,j)) + \tau$, for a given threshold $\tau > 0$. The definition of DCR is illustrated in Figure 1.1(a). Note that DCR intervals are defined *per pixel*, thus giving precise information about how much error can be tolerated in the disparity maps.

1.1.2 Application to temporal coding

The defined per-pixel DCRs give us a new degree of freedom in the encoding of disparity maps, where we are only required to reconstruct each depth pixel at the decoder to within its defined range of precision (as opposed to the original depth pixel), thus potentially resulting in further compression gain. Specifically, we change three aspects of the encoder in order to exploit DCRs: motion estimation, residual coding, and skip mode.

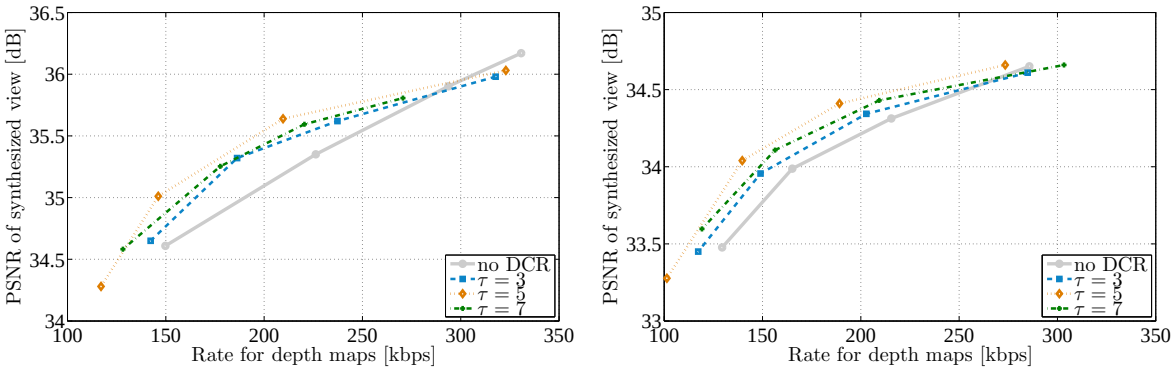
We start by defining a per-block *DCR space* for a target depth block $\mathcal{B}$ as the feasible space containing depth signals with each pixel falling inside its per-pixel DCR. As an example, Figure 1.1(b) illustrates the DCR space for a two-pixel block with per-pixel DCR $[2,6]$ and $[1,4]$. For a given predictor block, in order to minimize the energy of the prediction residuals, we *project the predictor on the DCR space*. In Figure 1.1(b), if the predictor is $(5,5)$, we identify $(5,4)$ in DCR space as the closest signal in DCR space, with resulting residuals $(0,-1)$. If the preditor is $(5,3)$, we identify $(5,3)$ in DCR space as the closest signal with residuals $(0,0)$.

In mathematical terms, we compute a prediction residual $r'(i,j)$ for each pixel (i,j) given predictor pixel value $\mathcal{P}(i,j)$ and DCR $[\mathrm{DCR}_{low}(i,j), \mathrm{DCR}_{up}(i,j)]$ according to the following soft-thresholding function:

$$r'(i,j) = \begin{cases} \mathcal{P}(i,j) - \mathrm{DCR}_{up}(i,j) & \text{if } \mathcal{P}(i,j) > \mathrm{DCR}_{up}(i,j), \\ \mathcal{P}(i,j) - \mathrm{DCR}_{low}(i,j) & \text{if } \mathcal{P}(i,j) < \mathrm{DCR}_{low}(i,j), \\ 0 & \text{otherwise.} \end{cases} \tag{1.1.3}$$

We use these residuals to find the motion vectors in the rate-distortion optimization. These residuals are then transformed, quantized and entropy coded, following the standard coding pipeline.

For SKIP macroblocks, no residual signal is coded, and thus we can not guarantee that the synthesized view error is bounded. Thus, we (conservatively) prevent the SKIP mode to be selected from the encoder if *any* reconstructed pixel of that macroblock violates DCR.



(a) *Kendo* (b) *Balloons*

Figure 1.1.2: RD curves for *Kendo* and *Balloons*

1.1.3 Experimental results

We modified an H.264/AVC encoder (JM reference software v. 18.0) in order to include DCR in the motion prediction and coding of residuals. Our test material includes 100 frames of two multiview video sequences, *Kendo* and *Balloons* with spatial resolution of 1024×768 pixels and frame rate equal to 30 Hz. For both sequences we coded the disparity maps $\mathbf{d}_3$ and $\mathbf{d}_5$ of views 3 and 5 (with IPP...GOP structure), using either the original H.264/AVC encoder or the modified one. In the latter case, we computed per-pixel DCRs with three values of τ, namely $\tau = \{3, 5, 7\}$. Given the reconstructed disparities in both cases (with/without DCR), we synthesize view $\mathbf{v}_4$ using the uncompressed views $\mathbf{v}_3$ and $\mathbf{v}_5$ and the compressed depths $\tilde{\mathbf{d}}_3$ and $\tilde{\mathbf{d}}_5$. Finally, we evaluate the quality of the reconstructed view $\hat{\mathbf{v}}_4$ w.r.t. ground-truth center view $\mathbf{v}_4$.

The resulting rate-distortion curves are reported in Figure 1.1.2. For the *Kendo* sequence, using $\tau = 5$ we obtain an average gain in PSNR of 0.34 dB and an average rate saving of about 28.5%, measured through the Bjontegaard metric. Notice that the proposed method enables a significant amount of bit saving by reducing *selectively* the fidelity of the reconstructed depths where this is not bound to affect excessively the synthesized view. On the other had, to achieve an equivalent bitrate reduction, a conventional decoder should quantize prediction residuals much more aggressively, and the quantization error can affect *all* the synthesized pixels.

We observe that most of the rate savings are obtained through a more efficient use of SKIP mode (which increases by over 18% in our experiments), and by a more efficient prediction of motion and coding of residuals. Notice that in the current setting, we are not taking into account the effect of quantization error, which could make reconstructed values lie outside DCR. Also, we optimize residuals in the spatial domain, while in practice they are transform coded. Jointly optimizing motion vectors and transform coded residuals, as well as pushing the de-quantized and reconstructed values inside DCR is a more challenging problem, and has been not considered in this work.

1.2 Optimal tone mapping for HDR video compression

The Human Visual System (HVS) is able to perceive a wide range of colors and luminous intensities, as present in real life outdoor scenes, ranging from bright sunshine to dark shadows. However, current traditional imaging technologies cannot capture nor reproduce such a broad range of luminance. The objective of High Dynamic Range (HDR) imaging is to overcome these limitations, hence leading to more realistic videos and a greatly enhanced user experience.

Whereas conventional Standard Dynamic Range (SDR) video has luminance values typically ranging from 0.1 to 100 cd/m^2, HDR video can represent a substantially higher peak luminance up to 10000 cd/m^2, providing bright pictures and wide contrast that result in a viewing experience closer to reality. Given that it entails a significantly higher raw data rate, efficient representation and coding of HDR video is one of the important issues to be addressed. In addition, the data characteristics also differ when compared to conventional SDR video content, calling for new coding approaches.

HDR video coding has been matter of standardization in MPEG [106]. The proposed solutions were based on the state-of-the-art video coding standard, high efficient video coding (HEVC) [81], and focused on HDR video compression efficiency using an electro-optical transfer function (EOTF) with 10 bit-depth coding profile [37]. Essentially, the EOTF is applied to provide a perceptually uniform representation which allows reducing the number of bits required for coding. Among several proposals, two perceptually optimized transfer functions, hybrid log-gamma (HLG) [13] and perceptual quantizer (PQ) [120], have been recommended for HDR video coding. Both transfer functions map absolute luminance values to perceptual codewords and share similarities with the perceptually uniform (PU) encoding introduced by Aydin *et al.* in [5]. However, they are mainly addressing two different applications. On one hand, HLG aims at providing a backward-compatible representation with 10 bit-depth video devices, especially suited for TV broadcasting services. On the other hand, PQ focuses on high bit precision representation coding, *e.g.,* 10 or 12 bit-depth video representation, which is *not* backward-compatible with the currently available SDR devices.

Unlike EOTF-based HDR video compression solutions, where a fixed curve is used for converting each HDR frame to a reduced dynamic range representation, in this work we consider a content-adaptive *tone mapping operator* (TMO). Therefore, our solution takes statistical characteristics of the input HDR frame into account. An illustration of this coding scheme is given in Figure 1.2.1: HDR video frames are first tone mapped to an SDR representation, which is fed

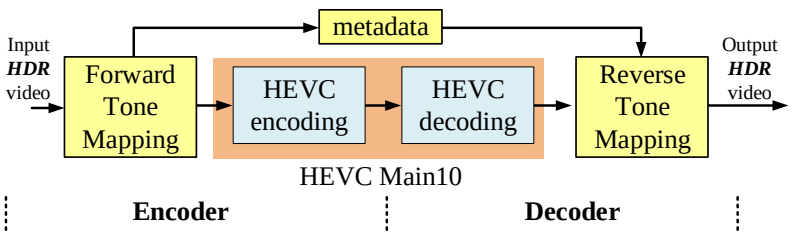


Figure 1.2.1: General diagram of the proposed HDR video coding method.

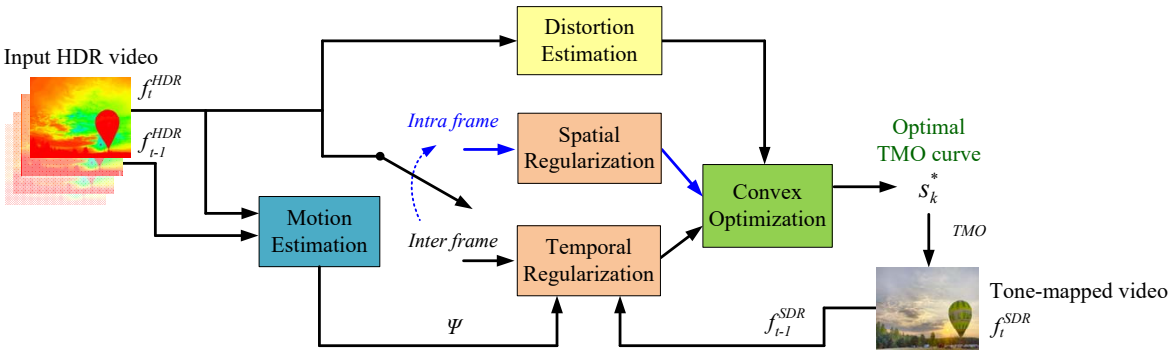


Figure 1.2.2: Block diagram of the proposed TMO. A content-adaptive spatially and temporally constrained tone mapping curve is obtained for each frame.

into a conventional video encoding pipeline. At the decoder side, the TMO is reversed (with losses), producing a reconstructed HDR video sequence. In addition to the SDR video bitstream, this scheme requires the transmission of *metadata*, e.g., the TMO parameters, in order to reverse the tone mapping operation at the decoder.

We employ the same piecewise-linear parametrization of a global tone mapping curve proposed by Mai *et al.* [107]. In that work the curve parameters are found by minimizing the mean squared error (MSE) between the original and the reconstructed HDR pixels. However, optimizing TMO *independently* for each frame does not take into account the spatial and temporal redundancies typical of a video signal. Instead, we introduce a content-adaptive, **spatio-temporal tone mapping operator (ST-TMO)** that includes spatial and temporal regularization terms, in order to find the optimal rate distortion (RD) tone-mapping curves for each frame. Through a comprehensive performance assessment including comparisons with state-of-the-art methods and multiple objective quality metrics, we show that the proposed scheme leads to significant coding gains over simpler tone mapping approaches and the MPEG PQ anchor.

The content of this section is described in greater detail in the papers [96, 137, 138].

1.2.1 Content Adaptive Tone Mapping Operator

The general scheme of the proposed ST-TMO is illustrated in Figure 1.2.2. For each input HDR video frame, we compute two cost terms: an estimation of the MSE between the original and the reconstructed HDR frame; and a regularization term, which enforces spatial or temporal coherence, depending on whether the frame is Intra or Inter predicted, respectively. Specifically, the temporal regularization relies on the knowledge of the motion field between the current and the previous frame (without loss of generality, we assume here that the temporal prediction at time t is obtained based on the frame at time $t-1$).

We express the unknown tone mapping curve to be found using the simple, piecewise-linear parametrization proposed in [107]. That is, the TMO is expressed as a vector s of slopes, as detailed below. This enables to define the cost terms mentioned above as convex functions of s, and to solve the resulting convex optimization problem through a proximal optimization

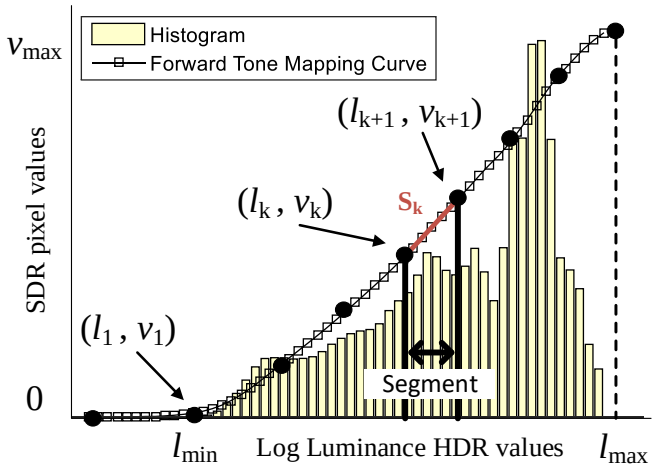


Figure 1.2.3: Piecewise parametrization of the optimal tone mapping curve as in [107].

method. As a result, we obtain a vector s^* of optimal slopes, which can be used to tone map the HDR picture into an SDR frame f_t^{SDR}. The vector s^* is sent as metadata information to the decoder in order to invert the tone mapping and reconstruct the HDR information. A more detailed description of the parametrization employed, the cost terms and the convex optimization follows.

TMO parametrization

The piecewise-linear TMO parametrization is illustrated in Figure 1.2.3. The histogram of the log-luminance l is divided into N segments of equal length δ. Let $k \in [1, \dots N]$ be the segment index, and p_k the associated probability mass. The TMO curve is then described by the set of nodes $\{l_k, v_k\}$, where v_k denotes the tone-mapped value; or, equivalently, by a vector $s \in \mathbb{R}_+^N$ composed by the non-negative slopes s_k for each segment, that is:

$$s_k = \frac{v_{k+1} - v_k}{\delta}. \tag{1.2.1}$$

Based on the s_k, the tone-mapped pixel values v are obtained as

$$v(l) = (l - l_k) \cdot s_k + v_k, \quad \forall l \in [l_k, l_k + 1), \tag{1.2.2}$$

where $l_1 = l_{\min}$ is the minimum luminance of the image. In order to inverse (1.2.2) and compute the reconstructed log-luminance, $\hat{l}$, the receiver needs to know as side information the slopes s_k, the value of δ as well as $l_{\min}$. Therefore, the size of the metadata depends on the number of segments, or equivalently on δ. Here we select $\delta = 0.1$ in $\log_{10}$ units to approximately match the Weber ratio (at least at high luminance), as proposed in [107]. For typical HDR content, this results in a number N of segments around $70 \sim 80$, which represents a modest overhead to overall transmission cost.

This parametrization can be used to find the optimal TMO curve, *i.e.,* the slopes s_k that minimize the squared reconstruction error $\varepsilon(s_k) = \|l - \hat{l}\|_2^2$. It is shown in [107] that, at least at high bitrates, the distortion is well approximated as a function of s_k as:

$$\varepsilon(s_k) \approx \sum_{k=1}^{N} \frac{p_k}{s_k^2}. \tag{1.2.3}$$

Based on this approximation, the authors of [107] find the slopes that minimize the squared-error distortion by formulating the following optimization problem:

$$\underset{s_k}{\text{minimize}}\ \varepsilon(s_k) \quad \text{subject to: } \sum_{k=1}^{N} s_k = \frac{v_{\max}}{\delta};\ s_k > 0, \tag{1.2.4}$$

where the constraint guarantees that the TMO curve spans all the available standard dynamic range (*e.g.,* $v_{\max} = 255$ for 8-bit images). This problem can be solved in closed form, yielding:

$$s_k = \frac{v_{\max} p_k^{1/3}}{\delta \sum_{n=1}^{N} p_n^{1/3}}. \tag{1.2.5}$$

Notice that this result is optimal in the MSE sense, but does not take into account the excess of bitrate produced by the possible loss of spatial and temporal coherency in the v signal. In order to take these effects into account, in the following we propose a spatial and a temporal regularization terms to be added to the problem in (1.2.4).

Spatial regularization

By simply optimizing (1.2.4), the resulting tone-mapped frames might lose spatial smoothness, which is typical in natural images, e.g., due to noise in the original HDR picture. This increases spatial complexity, and thus bitrate. To alleviate this, we modify problem (1.2.4) by adding a spatial regularization term, $C_{spa}(f_{intra}^{SDR})$, for tone-mapped Intra-coded images. C_{spa} is a real-valued convex function that models the spatial complexity, and f_{intra}^{SDR} is the intra frame of a given GOP of the SDR (*i.e.,* tone mapped) video. Since natural images usually exhibit a smooth spatial behavior, except around some locations (*e.g.,* object edges) where discontinuities arise, popular regularization models tend to penalize the image gradient. In this context, we adopt the total variation measure [18] due to its simplicity and effectiveness. Thus, we express the spatial regularization term as:

$$C_{spa}\left(f_{intra}^{SDR}\right) = \left\|\nabla f_{intra}^{SDR}\right\|_{1,2} = \sum_{i\in\Omega} \left\|(\nabla f_{intra}^{SDR})_i\right\|_2, \tag{1.2.6}$$

where Ω is the rectangular lattice over which the image f is defined, and $(\nabla f_{intra}^{SDR})_i$ is the 2-element vector denoting the gradient of f_{intra}^{SDR} at site i.

Notice that, since the log-luminance values l are constant for a given image, the tone mapping in (1.2.2) is linear in s, and can thus be conveniently rewritten as a matrix-vector multiplication

$f_{intra}^{SDR} = Zs$. Specifically, for a given HDR image with M pixels, $Z = [z_1, \ldots, z_M]^T$ is an $M \times N$ matrix, where each row has the form:

$$z_m = [\delta, \ldots, \delta, l - l_k, 0, \ldots, 0], \tag{1.2.7}$$

with the term $l - l_k$ in the k-th position if $l \in [l_k, l_{k+1})$. This formulation expresses the tone mapping equation (1.2.2), in that an HDR pixel l_i, falling in the j-th bin of the histogram, is mapped as $v_i = \sum_{k=1}^{j-1} \delta s_k + (l_i - l_j) s_j$.

Temporal regularization

For Inter-predicted tone-mapped video frames, applying the frame-by-frame tone mapping curve in (1.2.1) might lead to a loss of temporal coherence and a consequent increase of coding bitrate. We enforce temporal smoothness in the tone-mapped video by proposing a temporal regularization term, C_{temp}, to add to problem (1.2.4).

Specifically, C_{temp} is defined as:

$$C_{temp}(f_{inter}^{SDR}) = \sum_{i,j} \left(f_t^{SDR}(i,j) - \Phi \left[f_{t-1}^{SDR}(i,j); \Psi(i,j) \right] \right)^2 \tag{1.2.8}$$

where $\Phi[f_{t-1}(i,j); \Psi(i,j)] = f_{t-1}(i + \Psi(i,j)_x, j + \Psi(i,j)_y)$ is a function that gives the value of pixel at position (i,j) after motion compensation by the 2-element motion vector $\Psi(i,j)$ (x and y represent the horizontal and vertical components, respectively). The notation f_{inter}^{SDR} is used here to stress the fact that f_t is inter predicted. In order to get a precise per pixel motion field, we estimate Ψ by employing the optical flow algorithm in [21] directly on the original HDR frames f_t^{HDR} and f_{t-1}^{HDR}. This optical flow is then applied to obtain the motion compensated frame $\Phi \left[f_{t-1}^{SDR}(i,j); \Psi(i,j) \right]$.

Note that f_t^{SDR} is a function of s_k. By explicitly minimizing a temporal prediction residual, such a constraint leads to improved rate-distortion performance when encoding the tone mapped SDR video sequence. Instead of explicitly computing the sum of pixel-wise differences, as defined in Eq. (1.2.8), we notice that the temporal term C_{temp} is proportional to the expected temporal difference between two (motion-compensated) frames. By the definition of expected value, we can then compute (1.2.8) as the sum of all possible frame difference values, weighted by the probability of occurrence of each difference, that is:

$$C_{temp}(f_{inter}^{SDR}) = \sum_{k=1}^{N} \sum_{w=0}^{v_{max}} \left\{ \left(\frac{d_k}{2} + \sum_{i=1}^{k-1} d_i - w \right)^2 p_{k,w} \right\}, \tag{1.2.9}$$

where, for notational convenience, we define $d_k = \delta s_k$; $\frac{d_k}{2} + \sum_{i=1}^{k-1} d_i$ is then the SDR reconstruction value for pixels falling in the bin k of the log-luminance histogram as in Eq. (1.2.2), assuming a mid-tread quantizer on the real tone-mapped pixel values; $w \in [0, v_{max}]$ is the value of a pixel in the motion-compensated SDR frame $\Phi \left[f_{t-1}^{SDR}(i,j); \Psi(i,j) \right]$; and, finally, $p_{k,w} = \Pr\{f_t^{HDR} = l_k \wedge f_{t-1}^{SDR} = w\}$ is the joint probability that a pixel with log-luminance l_k in f_t^{HDR} has a motion-compensated predictor which has been tone mapped to the value w in f_{t-1}^{SDR}.

Notice that, while tone mapping f_t^{HDR}, the previous SDR frame f_{t-1}^{SDR} has been already computed, *i.e.,* w depends only on the (constant) motion vector field Ψ. In practice, we pre-compute $p_{k,w}$ after motion estimation, before computing Eq. (1.2.9). Finally, as shown more in detail in [138], the temporal constraint in (1.2.9) can be rewritten as a quadratic function:

$$C_{temp}(f_{inter}^{SDR}) = s^T W_2 s + s^T W_1, \tag{1.2.10}$$

where W_1 and W_2 are constant matrices defined in [138].

Convex Optimization

Based on the spatial and temporal constraints defined above, we can redefine the optimization problem for each frame f_t^{HDR} in Eq. (1.2.4) as:

$$\underset{s_k}{\text{minimize}}\ \hat{\varepsilon}(s_k) + \mathcal{C} \quad \text{subject to: } \sum_{k=1}^{N} s_k = \frac{v_{\max}}{\delta}, \tag{1.2.11}$$

where the term:

$$\mathcal{C} = \begin{cases} \lambda_{spa} C_{spa}(f_{intra}^{SDR}) & \text{if } f_t^{HDR} \text{ is Intra-predicted} \\ \lambda_{temp} C_{temp}(f_{inter}^{SDR}) & \text{if } f_t^{HDR} \text{ is Inter-predicted.} \end{cases}$$

The weighting terms λ_{spa} and λ_{temp} define the relative importance of spatial/temporal smoothness with respect to MSE minimization, and the details about how to determine them are given in [138]. Notice that Problem (1.2.11) consists in minimizing the sum of two convex functions over the convex set:

$$\Theta = \left\{ s \in \mathbb{R}^N \;\middle|\; \sum_{k=1}^{N} s_k = \frac{v_{max}}{\delta} \right\}. \tag{1.2.12}$$

Moreover, the term $\hat{\varepsilon}(s_k)$, defined as:

$$\hat{\varepsilon}(s_k) = \begin{cases} +\infty & \text{if } s_k \leq 0 \\ \varepsilon(s_k) & \text{otherwise,} \end{cases} \tag{1.2.13}$$

is also convex.

In order to solve Problem (1.2.11), we employ here a proximal algorithm [23], which enables to address both non-smooth functions and hard constraints, without requiring any matrix inversion.

1.2.2 Experimental results

We evaluate the compression performance of the proposed ST-TMO through an extensive set of experiments. The input to ST-TMO is an HDR video frame in linear (photometric) RGB domain. The required color conversion and chroma sub-sampling algorithms suggested in the MPEG CfE on HDR and WCG video coding [106] are used. We employ the optical flow

algorithm (with the publicly available implementation) in [43, 21] (details about the optical flow configuration are reported in [138]).

Experiments were carried out for eight high-definition (HD) test sequences, namely: *SunRise*, *Smith_Welding*, *EBU_06_Starting*, *Market3Clip4000r2*, *FireEater2Clip4000r1*, *EBU_04_Hurdles*, *Carousel_fireworks_03*, and *Carousel_fireworks_04*. These sequences have been selected in order to provide a varied sample of spatial/temporal complexity and dynamic range.

We compare the proposed ST-TMO with the frame-by-frame TMO of Mai *et al.* (2011) [107] and with the anchor solution based on the PQ EOTF proposed in the MPEG CfE [106], available in the MPEG HDRTools v.0.17 [187]. In addition, we implemented the temporally optimized TMO of Mai *et al.* (2013) [108], where flickering is reduced by penalizing the difference between the average brightness of consecutive tone-mapped frames, instead of local (per pixel) motion trajectories as in ST-TMO. The rate overhead from the all the tone mapping methods (*i.e.,* transmitting metadata) is *included* in all reported results, and is around 130 kbps on average for the considered sequences.

In order to evaluate the compression performance, video test sequences were encoded using the HEVC reference model (HM) ver. 16.2 software. The low-delay HEVC encoder configuration we used is as follows: GOP length of 16, predictive coded (P) pictures, YCbCr 4:2:0 chroma sampling, and an internal bit-depth of 10. We set v_{max} to 1023 for TMOs. Variation in bitrates was achieved using different quantization parameter (QP) values.

Evaluation of HDR video distortion is more challenging than conventional SDR quality assessment (see Section 2.1). In order to provide a more informative comparison, we evaluate HDR video distortion using several metrics: Peak signal to noise ratio (PSNR) and structural similarity index (SSIM) [196], both computed on either log-luminance or PU-encoded values [5]; HDR-Visible differences predictor (HDR VDP 2.2.1) [131]; the HDR-Video Quality Measure (HDR-VQM) [129]; HDR Metrics in the MPEG HDRTools [187], including tPSNR, L_{100} and L_{1000}.

We report average Bjøntegaard delta (BD) metric [11] gains in Table 1.2.1. These values are the average gain in the corresponding quality metric. We opted for reporting BD quality instead of BD rate gains since the computation of the latter is unstable when the RD curves are not properly aligned on the quality axis From the table, it is evident that the proposed ST-TMO leads to superior coding performance in most cases, and on average yields consistent gains with all the considered quality metrics with respect to both a fixed transfer function and a state-of-the-art TMO-based HDR video compression scheme. In particular, our proposed TMO is beneficial for contents that display local motion and high spatial and temporal complexity.

It must be mentioned that the proposed ST-TMO has a higher complexity than a fixed EOTF or a frame-by-frame TMO (see [138] for a more detailed analysis). In fact, if a per pixel motion estimation enables to achieve higher coding gain compared to [108], the optical flow is responsible for a large increase in the computation time. Thus, an interesting follow-up direction would be to study how the performance of ST-TMO are affected by coarser and faster motion estimation techniques. In addition, the current optimization employs mean squared error as a fidelity criterion, as this brings to a convex problem formulation. It might be interesting to study how more perceptually motivated fidelity criteria (e.g., based on structural similarity) could be employed

Table 1.2.1: Quality gain of the proposed ST-TMO in terms of BD quality (dB) gains. The highest BD-quality gains in **blue** and BD-quality losses in *red*.

Method	Sequence	log-PSNR	log-SSIM	PU-PSNR	PU-SSIM	HDR-VDP Q	HDR-VQM	tPSNR	L_{100}	L_{1000}
PQ	*Market3Clip4000r2*	0.247	0.001	*-0.070*	*-0.001*	1.029	0.071	0.252	*-0.241*	*-0.272*
Mai *et al.* (2011)		**1.507**	**0.011**	**0.85**	0.002	**2.119**	**0.185**	**1.468**	**0.237**	**0.216**
Mai *et al.* (2013)		0.813	0.003	0.543	**0.002**	1.198	0.070	0.687	0.066	0.061
PQ	*FireEater2Clip4000r1*	0.794	0.008	0.002	*-0.001*	*-0.367*	0.056	4.469	1.911	1.903
Mai *et al.* (2011)		**4.556**	**0.028**	**3.370**	**0.014**	*-0.309*	0.074	5.863	**2.299**	**2.104**
Mai *et al.* (2013)		0.221	*-0.001*	0.789	0.001	**0.579**	**0.088**	**10.180**	*-1.219*	*-0.985*
PQ	*SunRise*	**5.337**	0.004	**6.838**	**0.003**	**7.265**	0.020	**4.537**	3.492	0.060
Mai *et al.* (2011)		3.516	**0.006**	4.794	0.002	2.902	**0.030**	0.071	**3.867**	**3.932**
Mai *et al.* (2013)		*-0.896*	*-0.003*	*-0.892*	*-0.002*	0.638	*-0.030*	*-0.477*	*-0.154*	*-0.155*
PQ	*EBU_04_Hurdles*	0.128	*-0.006*	*-0.217*	*-0.008*	5.324	0.142	6.402	0.915	0.932
Mai *et al.* (2011)		1.477	*-0.001*	1.239	*-0.001*	2.901	0.042	2.138	**1.351**	**1.357**
Mai *et al.* (2013)		**2.514**	**0.017**	**2.437**	**0.021**	**11.581**	**0.240**	**6.685**	1.202	1.196
PQ	*EBU_06_Starting*	0.959	0.001	0.217	*-0.001*	**2.744**	0.017	**2.276**	1.358	1.369
Mai *et al.* (2011)		**2.025**	**0.006**	**1.595**	0.002	2.226	**0.095**	2.702	**1.389**	**1.401**
Mai *et al.* (2013)		0.952	0.003	0.744	**0.003**	1.716	0.011	1.485	0.912	0.919
PQ	*Carousel_fireworks_03*	3.688	0.042	4.770	0.008	2.172	0.158	3.598	1.288	1.363
Mai *et al.* (2011)		**7.529**	**0.147**	**10.294**	**0.040**	**4.323**	0.065	**7.438**	**4.209**	**4.798**
Mai *et al.* (2013)		1.249	0.020	1.441	0.01	1.670	**0.280**	1.837	0.865	0.915
PQ	*Carousel_fireworks_04*	4.819	0.021	5.679	0.013	4.523	**0.247**	5.543	1.558	1.465
Mai *et al.* (2011)		**7.862**	0.043	**9.074**	0.023	**5.254**	*-0.014*	**7.987**	**3.626**	**3.635**
Mai *et al.* (2013)		2.938	**0.052**	3.649	**0.040**	3.349	0.203	2.798	1.698	1.640
PQ	*Smith_Welding*	0.071	0.014	*-1.091*	*-0.001*	1.406	*-0.138*	*-0.335*	0.558	0.752
Mai *et al.* (2011)		**14.954**	**0.909**	**14.077**	**0.093**	**3.466**	0.024	**2.057**	**7.232**	**6.778**
Mai *et al.* (2013)		0.366	0.002	1.121	0.001	0.031	**0.072**	1.032	0.432	0.698
				Average						
PQ		2.005	0.011	2.018	0.002	**3.012**	0.072	3.343	1.355	0.946
Mai *et al.* (2011)		**5.428**	**0.144**	**5.662**	**0.022**	2.860	0.063	**3.716**	**3.026**	**3.027**
Mai *et al.* (2013)		1.02	0.012	1.229	0.008	2.595	**0.117**	3.028	0.475	0.536

to further increase coding gains.

1.3 Enhancing Intra prediction using context-based learning

This section describes ongoing work on using machine learning tools (and specifically, deep convolutional neural networks) to enhance parts of the state-of-the-art video coding pipeline, such as spatial/temporal prediction. In particular, the work on Intra prediction reported in the following has been recently published in [194].

Modern image and video codecs strongly rely on spatial prediction as a fundamental tool to achieve high rate-distortion performance. Conventionally, spatial prediction leverages an ensemble of simple linear models to interpolate information from a context of already decoded pixels, with the goal to obtain a prediction residual which is simpler to code than the original signal. These prediction models have been improved and optimized for several decades, by continuously adding new modes, block partitions and prediction directions, e.g., 33 directional modes and up to 32×32 prediction units are employed in the HEVC video standard [180].

Despite the high number of available prediction modes, current spatial prediction approaches assume the underlying signal can be approximated by a simple linear combination of a few (reconstructed) pixels. Increasing further the number of prediction modes might guarantee a better signal approximation; however, this leads to continuously increase the computational

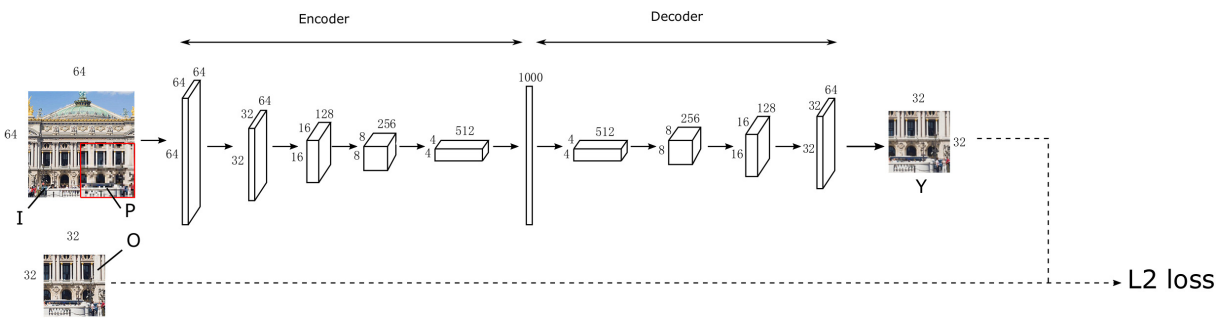


Figure 1.3.1: The proposed encoder-decoder convolutional network architecture.

cost, and is still ineffective when the signal to predict requires more complex representations than simple bilinear interpolation.

Recently, deep representation learning models such as auto-encoders have been employed to learn effective representations for very complex signals such as natural images [42]. Differently from the signal models used in video spatial prediction, deep auto-encoders are much more complex and highly non-linear. In the last couple of years, these models have been applied to image compression [182, 1, 121, 162, 6, 167], yielding in many cases equivalent or better visual quality than traditional image codecs [192].

In this work we aim at improving the spatial prediction produced by state-of-the-art codecs such as HEVC, in order to reduce the bitrate of the prediction residual. Differently from the work cited above on learning-based image coding, we do not propose here an end-to-end coding approach, but rather to improve the already well-optimized spatial prediction of HEVC by leveraging the expresiveness of a deep convolutional representation. To this end, we employ a convolutional encoder-decoder neural network to predict a block of an image based on a context of already decoded pixels *and* the spatial prediction of the same block as *produced by the rate-distortion optimization* of the video encoder. We refer to the proposed scheme as *context-based prediction enhancement* (CBPE). Our experiments on a dataset of natural images show that the proposed CBPE reduces the mean square error of HEVC spatial prediction by 25% on average.

1.3.1 Network

In order to improve the HEVC Intra predictor of HEVC we propose the CBPE network architecture shown in Fig. 1.3.1. It is composed of an *encoder* connected to a *decoder* via a *bottleneck* [139]. The encoder takes in input a 64×64 image I and projects it on a latent feature space. The decoder recovers from the features produced by the bottleneck layer a 32×32 image.

Our network is trained end-to-end to reconstruct an original (uncompressed) 32×32 block O starting from an input 64×64 image I, as shown in Figure 1.3.1. The bottom-right quadrant of the input block is the HEVC prediction $\mathcal{P}$ obtained after rate-distortion optimization, while the remaining quadrants are the decoded (thus, noisy) causal context of the block to be predicted.

The loss function is defined as the mean squared error (MSE) between $\mathcal{Y}$ and O, that is:

$$L(w,y,o) = \frac{1}{K}\sum_k (y_k(w) - o_k)^2, \tag{1.3.1}$$

where y_k and o_k are pixels of $\mathcal{Y}$ and O, respectively, and $K = 32^2$. Further details about the network architecture, initialization and optimization parameters are available in [194].

1.3.2 Experimental results

The training and validation of the proposed CBPE is carried out by drawing at random about 16k images from the dataset originally proposed in [168]. The dataset contains natural images downloaded from Flickr, spanning a wide range of semantic classes and acquisition quality. The images come in a JPEG compressed format, with the original quality/resolution of the Flickr source, thus providing a large variety of train/test conditions. Each image is independently compressed and decoded using the H.265/HEVC HM reference software (version 16.0), with a fixed $QP = 15$. All prediction unit sizes, from 4×4 to 32×32 are enabled in the rate-distortion optimization. Next, for each decoded image, we extract a number of non-overlapping 64×64 patches aligned with the HEVC CTU grid, along with the HEVC predictors $\mathcal{P}$. Following this protocol, a total of 405k patches are extracted from a first set of randomly drawn images, of which 324k (80%) are used for training and 81k (20%) for validation. Finally, about 50k patches are extracted from a different set of randomly drawn images for testing.

In order to test the performance of the proposed CBPE, we provide in input to our trained network the test patches and, for each test patch, we measure the MSE between the network output $\mathcal{Y}$ and the ground-truth, uncompressed reference O, i.e., the energy of the prediction residual obtained by CBPE. For comparison, we also compute for each test patch the energy of the HEVC prediction residual, i.e., the MSE between HEVC predictor $\mathcal{P}$ and O. In approximately 66% of the cases the CBPE *enhances* the HEVC predictor by reducing the energy of the prediction residuals. The MSE of CBPE reduces the average HEVC predictor MSE by about 25%. Although quantifying precisely the end-to-end coding gain provided by CBPE would require integrating it into a whole coding chain, we notice that reducing the energy of prediction residuals can generally improve rate-distortion performance in predictive coding.

In order to illustrate qualitatively the prediction improvement brought by CBPE, we report in Figure 1.3.2 a few examples of predicted blocks. From left to the right, we show the original content, the HEVC predictor, the predictor refined by CBPE, the HEVC prediction residual and the residual after CBPE, for three different patches, with the corresponding prediction MSE. We observe that, in these cases, the HEVC predictor can capture the overall structure of the block. However, due to the limited directional prediction modes and the block-based predictions, the HEVC prediction alone introduces some visible artifacts, and fails in capturing fine-grained structures of the content. Conversely, the CBPE can enhance this prediction, smoothing out the HEVC blocking and recovering somehow better the original image structure. Interestingly, the CBPE predictor has a more natural aspect, confirming previous findings on the ability of deep generative models to learn image "naturalness" [192].

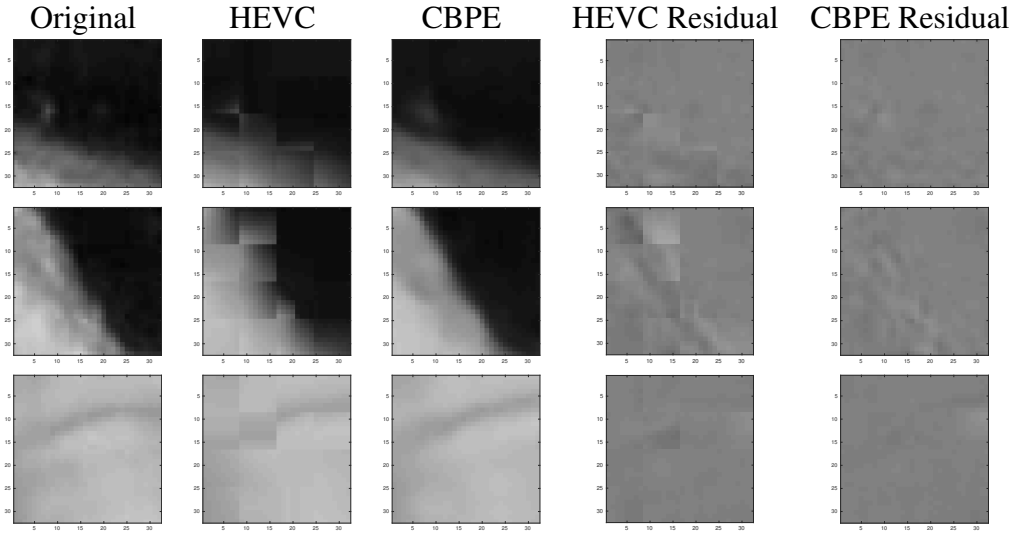


Figure 1.3.2: Comparison of predicted patches by using HEVC prediction and proposed scheme prediction methods. The MSE of prediction blocks using HEVC and CBPE for each blocks are: top: 83.99 (HEVC), 34.99 (CBPE); middle: 321.71 (HEVC), 57.37 (CBPE); bottom: 36.63 (HEVC), 27.41 (CBPE).

It is also instructive to analyze cases where CBPE fails and degrades the quality the HEVC predictor. Figure 1.3.3 shows two examples. We observe that CBPE tends to over-smooth patches with periodic or high-frequency structures and sharp edges, and in some cases to add some low-frequency noise which was not present in the original signal. This might be caused by the lack of sufficient training data. An interesting solution to explore would be to add a regularization term in the loss function (1.3.1) in order to preserve sharper structures and penalize noise, similar to what is done in total-variation denoising.

This work provides a proof of concept that deep representations have a large potential to be used in video compression, by extending or replacing the conventional linear prediction and transform tools. However, how to optimally employ a deep predictor in a state-of-the-art video codec still needs a substantial amount of further research. From a practical point of view, an implementation of this work in a video codec would require a network for each possible prediction unit size, due to the progressive nature of spatial prediction, in order to maintain the synchronization between encoder and decoder. On the other hand, from a more speculative point of view, we are currently considering how prediction performance might be further enhanced by classifying (in an unsupervised way) the content of a block prior to CBPE.

1.4 Learning-based coding of point cloud geometry

This section describes preliminary work on coding of point cloud video. Point clouds are an essential data structure for Virtual Reality (VR) and Mixed Reality (MR) applications. A point cloud is a set of points in the 3D space represented by coordinates x, y, z and optional attributes (for example color, normals, etc.). Point cloud data is often very large as point clouds easily

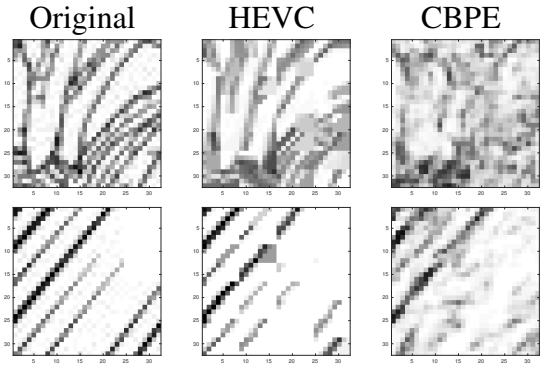


Figure 1.3.3: Some failure cases where CBPE is not able to improve HEVC spatial prediction. The MSE of prediction blocks using HEVC and CBPE for each blocks are: top: 2.25e+03 (HEVC), 3.53e+03 (CBPE); bottom: 1.45e+03(HEVC), 2.54e+03 (CBPE).

range in the millions of points and can have complex sets of attributes. Therefore, efficient point cloud compression (PCC) is particularly important to enable practical usage in VR and MR applications.

The Moving Picture Experts Group (MPEG) is currently working on PCC. In 2017, MPEG issued a call for proposals (CfP) and in order to provide a baseline, a point cloud codec for tele-immersive video [115] was chosen as the MPEG anchor. Research on PCC can be categorized along two dimensions. On one hand, one can either compress point cloud geometry, i.e., the spatial position of the points, or their associated attributes. On the other hand, we can also separate works focusing on compression of dynamic point clouds, which contain temporal information, and static point clouds.

In this work, we focus on the lossy compression of static point cloud geometry. In PCC, a precise reconstruction of geometric information is of paramount importance to enable high-quality rendering and interactive applications. For this reasons, lossless geometry coding has been investigated recently in MPEG, but even state-of-the-art techniques struggle to compress beyond about 2 bits per occupied voxels (bpov) [40]. This results in large storage and transmission costs for rich point clouds.

Lossy compression proposed in the literature, on the other hand, are based on octrees which achieve variable-rate geometry compression by changing the octree depth. Unfortunately, lowering the depth reduces the number of points exponentially. As a result, octree based lossy compression tends to produce “blocky” results at the rendering stage with medium to low bitrates. In order to partially attenuate this issue, [89] proposes to use wavelet transforms and volumetric functions to compact the energy of the point cloud signal. However, since they still employ an octree representation, their method exhibits rapid geometry degradation at lower bitrates. While previous approaches use hand-crafted transforms, we propose here a data driven approach based on learned convolutional transforms which directly works on voxels.

Specifically, we present a method for learning analysis and synthesis transforms suitable for point cloud geometry compression. In addition, by interpreting the point cloud geometry as a binary signal defined over the voxel grid, we cast decoding as the problem of classifying whether a given voxel is occupied or not. We train our model on the ModelNet40 mesh dataset

[202, 169], test its performance on the Microsoft Voxelized Upper Bodies (MVUB) dataset [103] and compare it with the MPEG anchor [115]. We find that our method outperforms the anchor on all sequences at all bitrates. Additionally, in contrast to octree-based methods, ours does not exhibit exponential diminution in the number of points when lowering the bitrate. We also show that our model generalizes well by using completely different datasets for training and testing.

This work is described in further details in the paper [150].

1.4.1 Proposed PCC coding scheme

We define the set of possible points at resolution r as $\Omega_r = [0 \dots r]^3$. Then, we define a point cloud as a set of points $S \subseteq \Omega_r$ and its corresponding voxel grid v_S as the following binary occupancy map:

$$\begin{aligned} v_S \colon \Omega_r &\longrightarrow \{0,1\}, \\ z &\longmapsto \begin{cases} 1, & \text{if } z \in S \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

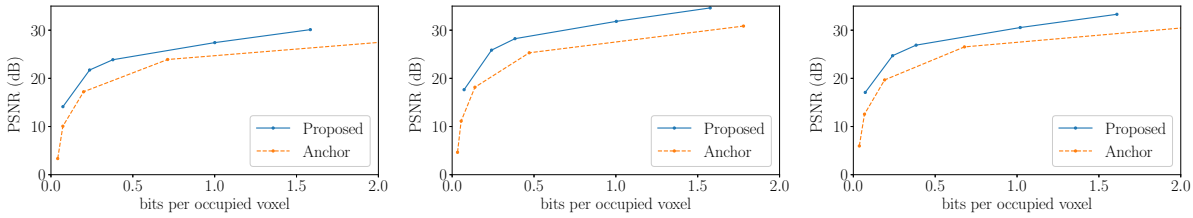
We use a 3D convolutional auto-encoder composed of an analysis transform f_a, followed by a uniform quantizer and a synthesis transform f_s. Let $x = v_S$ be the original point cloud. The corresponding latent representation is $y = f_a(x)$. To quantize y, we introduce a quantization function Q so that $\hat{y} = Q(y)$. This allows us to express the decompressed point cloud as $\hat{x} = \hat{v}_S = f_s(\hat{y})$. Finally, we obtain the decompressed point cloud $\tilde{x} = \tilde{v}_S = round(\min(0, \max(1, \hat{x})))$ using element-wise minimum, maximum and rounding functions. The functions f_a and f_s are convolutional neural networks. The details about the architecture of these networks, as well as the definition of 3D convolution and deconvolution operations used in this work, are available in [150]. We use the Adam optimizer [82] to learn the weights for our auto-encoder.

We handle quantization similarly to [6]. Q represents element-wise integer rounding during evaluation and Q adds uniform noise between -0.5 and 0.5 to each element during training which allows for differentiability. To compress $Q(y)$, we perform range coding and use the Deflate algorithm, a combination of LZ77 and Huffman coding [47] with shape information on x and y added before compression. Note however that our method does not assume any specific entropy coding algorithm.

Our decoding process can also be interpreted as a binary classification problem where each point $z \in \Omega_r$ of the voxel grid is either present or not. This allows us to decompose $\hat{x} = \hat{v}_S$ into its individuals voxels z whose associated probability of occupancy is p_z. However, as point clouds are usually very sparse, most $v_S(z)$ values are equal to zero. To compensate for the imbalance between empty and occupied voxels we use the α-balanced focal loss as defined in [101]:

$$FL(p_z^t) = -\alpha_z (1 - p_z^t)^\gamma log(p_z^t) \tag{1.4.1}$$

with p_z^t defined as p_z if $v_S(z) = 1$ and $1 - p_z$ otherwise. Analogously, α_z is defined as α when



(a) Andrew (−47.8% BDBR) (b) David (−55.7% BDBR) (c) Phil (−49.0% BDBR)

Figure 1.4.1: RD curves for three sequences of the MVUB dataset. We compare our method to the MPEG anchor.

$v_S(z) = 1$ and $1-\alpha$ otherwise. The focal loss for the decompressed point cloud can then be computed as follows:

$$FL(\tilde{x}) = \sum_{z \in S} FL(p_z^t). \tag{1.4.2}$$

Our final loss is $L = \lambda D + R$ where D is the distortion calculated using the focal loss and R is the rate in number of bits per input occupied voxel (bpov).

Notice that the rate is computed differently during training and during evaluation. On one hand, during evaluation, as the data is quantized, we compute the rate using the number of bits of the final compressed representation. On the other hand, during training, we add uniform noise in place of discretization to allow for differentiation. It follows that the probability distribution of the latent space $Q(y)$ during training is a continuous relaxation of the probability distribution of $Q(y)$ during evaluation which is discrete. As a result, entropies computed during training are actually differential entropies, or continuous entropies, while entropies computed during evaluation are discrete entropies. During training, we use differential entropy as an approximation of discrete entropy. This makes the loss differentiable which is essential for training neural networks.

1.4.2 Experimental results

We train and evaluate our network on the ModelNet40 aligned dataset [202]. The ModelNet40 dataset contains 12,311 mesh models from 40 categories. This dataset provides us with both variety and quantity to ensure good generalization when training our network. To convert this dataset to a point cloud dataset, we first perform sampling on the surface of each mesh. Then, we translate and scale it into a voxel grid of resolution r. We use this dataset for training with a resolution $r = 64$. Details about the training parameters (learning rate, batch size, etc.) are reported in [150].

Then, we perform tests on the MVUB dataset and we compare our method with the MPEG anchor [115]. The MVUB dataset [103] contains 5 sequences captured at 30 fps during 7 to 10 seconds each with a total of 1202 frames. We test our method on each individual frame with a resolution $r = 512$. In other words, we evaluate performance for intra-frame compression on each sequence.

(a) Original (b) Proposed method (c) MPEG Anchor

Figure 1.4.2: Original point cloud (left), the compressed point cloud using the proposed method (middle) and the MPEG anchor (right). Colors are mapped using nearest neighbor matching. Our compressed point cloud was compressed using $\lambda = 10^{-6}$ with a PSNR of 29.22 dB and 0.071 bpov. The anchor compressed point cloud was compressed using a depth 6 octree with a PSNR of 23.98 dB and 0.058 bpov. They respectively have 370,798; 1,302,027; and 5,963 points.

We compute RD curves for each sequence of the test dataset. For our method, we use the following λ values to compute RD points : 10^{-4}, 5×10^{-5}, 10^{-5}, 5×10^{-6} and 10^{-6}. For each sequence, we average distortions and bitrates over time for each λ to obtain RD points. For the MPEG anchor, we use the same process with different octree depths.

To evaluate distortion, we use the *point-to-plane symmetric PSNR* [184], that is, $e_{symm}(A,B) = \min(e(A,B), e(B,A))$ where $e(A,B)$ provides the point-to-plane PSNR between points in A and their nearest neighbors in B. This choice is due to the fact that original and reconstructed point clouds may have a very different number of points, e.g., in octree-based methods the compressed point cloud has significantly less points than the original, while in our method it is the opposite.

Our method outperforms the MPEG anchor on all sequences at all bitrates. The latter has a mean bitrate of 0.719 bpov and a mean PSNR of 16.68 dB while our method has a mean bitrate of 0.691 and a mean PSNR of 24.11 dB. RD curves and the Bjontegaard-delta bitrates (BDBR) for three sequences are reported in Figure 1.4.1. Our method achieves 51.5% BDBR savings on average compared to the anchor.

In Figure 1.4.2, we show examples on the first frame of the Phil sequence. Our method achieves lower distortion at similar bitrates and produces more points than the anchor which increases quality at low bitrates while avoiding “blocking” effects. This particular example shows that our method produces 218 times more points than the anchor at similar bitrates. In other words, both methods introduce different types of distortions. Indeed, the number of points produced by octree structures diminishes exponentially when reducing the octree depth. Conversely, our method produces more points at lower bitrates as the focal loss penalizes false negatives more heavily.

We are currently extending this work towards the compression of point cloud attributes. In particular, we are exploring the use of graph convolutions to compress attributes given the graph geometry.

Chapter 2

Quality of Experience

Assessing visual Quality of Experience is of paramount importance for the evaluation, tuning and design of image/video processing and compression pipelines. This chapter provides a summary of some contributions in QoE assessment for image and video. These contributions can be grouped along two main axes: on one hand, research work aimed at predicting visual quality or perceptual attributes through objective measures, in particular for emerging video formats such as high dynamic range imaging; on the other hand, research focusing on subjective methodologies for collecting ground-truth quality scores.

In Section 2.1, I describe an extensive evaluation campaign of full-reference quality metrics for HDR images and video [210]: in addition to the standard statistical performance evaluation, we also propose a new discriminability analysis methodology which takes into account the stochastic nature of ground-truth mean opinion scores. The second contribution (Section 2.2 is a no-reference quality assessment method for HDR pictures [84, 84]. The originality of the work consists in disentangling perceptual quality as a combination of two distinct processes: the physical (per pixel) error, and the perceptual scaling of this error that leads to the overall quality judgment. We model these two terms using convolutional neural networks. As a third contribution, I summarize in Section 2.3 a series of studies we conducted with the goal of measuring the perceived dynamic range (PDR) of an HDR picture [48, 52, 49]. In contrast with conventional psychophysical experiments on simple stimuli, assessing the perceived dynamic range on complex stimuli requires designing a new subjective methodology that takes into account the constraints of content visualization and the definition of the perceptual attribute. The collected data is used to fit a simple linear model to predict PDR based on image features.

Finally, the fourth contribution described in Section 2.4 [142, 207, 211] aims at exploring the relation between quality scales induced by rating and pairwise comparisons experiments. We recall the principles of scaling preferences onto an interval scale, showing the importance of comparing pairs of stimuli coming from *different* contents. These results are used later to build a model to merge rating and pairwise comparison scores.

2.1 An extensive evaluation of HDR fidelity metrics

As introduced in Section 1.2, evaluating High Dynamic Range (HDR) visual quality presents new challenges compared to conventional Low Dynamic Range (LDR) images [134]. The higher peak brightness and contrast offered by HDR increases the visibility of artifacts, and at the same time changes the way viewers focus their attention compared to LDR [133]. Since these and other factors intervene in a complex way to determine HDR visual quality, the most accurate approach to assess it is, in general, through subjective test experiments. However, these are expensive to design and implement, require specialized expertize and are time-consuming. Furthermore, in the case of HDR, subjective testing requires specialized devices such as HDR displays, which still have a high cost and a limited diffusion. Therefore, designing and tuning *full-reference* (fidelity) quality metrics for HDR content has been an important research topic [110, 129, 131, 5].

Two main approaches have been proposed to measure HDR fidelity. On one hand, some metrics require modeling of the human visual system (HVS), such as the HDR-VDP [110]. On the other hand, one can resort to metrics developed in the context of LDR imagery, such as PSNR or SSIM [195]. All these LDR metrics are based on the assumption that pixel values are perceptually linear, i.e., equal increments of pixel values correspond to equivalent changes in the perceived luminance. This is not true in the case of HDR content, where pixel values store *linear* light, i.e., pixels are proportional to the physical luminance of the scene. Instead, human perception has a more complex behavior: it can be approximated by a square-root in low luminance values and is approximately proportional to luminance ratios in higher luminance values, as expressed by the De Vries-Rose and Weber-Fechner laws, respectively [92]. Thus, in order to employ these metrics, the HDR content needs to be perceptually linearized, e.g., using a logarithmic or perceptually uniform (PU) encoding [5].

Previous studies about the performance of quality metrics show sometimes discrepancies in their conclusions about the ability of these metrics to yield consistent and accurate predictions of ground-truth Mean Opinion Scores (MOS). The aim of this work is to provide an extensive, reliable, and consistent benchmark of the most popular HDR image fidelity metrics. To this end, we collected as many as possible publicly available databases of HDR compressed images with subjective scores, in addition to proposing a new one which mixes different codecs and pixel encoding functions. This gives a total of 690 HDR images, which is by far *larger* than previous studies on HDR visual quality. We then align the MOS's of these databases using the iterated nested least square algorithm (INLSA) proposed in [146], in order to obtain a common subjective scale. Based on this data, we analyze the prediction accuracy and the discriminability (i.e., the ability of detecting when two images have different perceived quality) of 25 fidelity metrics, including those employed in MPEG standardization.

The content of this Section is described in greater detail in the original paper [210].

Table 2.1.1: Number of observers, subjective methodology, number of stimuli, compression type and tone mappings employed in the HDR image quality databases used in this section. TMOs legend: *AS*: Ashikmin, *RG*: Reinhard Global, *RL*: Reinhard Local, *DR*: Durand, *Log*: Logarithmic, *MT*: Mantiuk.

No	Obs.	Meth.	Stim.	Compr.	TMO
#1 [132]	27	ACR-HR	140	JPEG	iCAM [90]
#2 [130]	29	ACR-HR	210	JPEG 2000	AS [4] RG [158] RL [158] DR [32] Log
#3 [83]	24	DSIS	240	JPEG-XT	RG [158] MT [111]
#4 [189]	15	DSIS	50	JPEG JPEG 2000 JPEG-XT	Mai [107]
#5	15	DSIS	50	JPEG JPEG 2000	Mai [107] PQ [120]

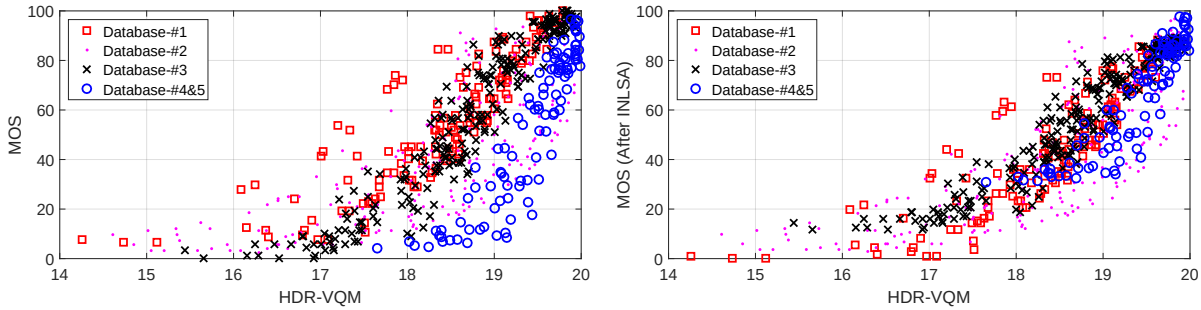
2.1.1 Datasets and quality metrics

Subjective datasets

In order to provide a solid evaluation of objective quality metrics, we considered 5 subjectively annotated datasets. The stimuli in each dataset have been obtained by compressing original HDR pictures (represented on floating point values). We focus on coding distortion as this is a typical test condition for fidelity metrics, and for the relevance this topic has had in the context of HDR video standardization. Specifically, except the JPEG-XT standard [161], the remaining compression schemes employ a tone mapping operator (TMO) to convert the HDR picture into a conventional, 8-bit representation, in order to use a standard image codec such as JPEG or JPEG2000, similar to the approach described in Section 1.2. The TMO is reversed at the decoder to reproduce a reconstructed HDR image. The key features of these datasets are summarized in Table 2.1.1. Notice that Dataset #4 was published in our previous work [189], while Dataset #5 is one of the contributions of this work. Further details about the selected datasets are reported in [210].

Alignment of the subjective scores

The datasets reported in Table 2.1.1 have been collected with different subjective methodologies and in different experimental conditions. Therefore, the corresponding mean opinion scores are not lying, in general, on the same scale. In Fig. 2.1.1(a), we observe the MOS distribution for non-aligned databases as a function of the HDR-VQM metric. We notice that different datasets tend to have a different relationship between the objective metric and the subjective scores. In



(a) Before INLSA (RMSE = 44.04) (b) After INLSA (RMSE = 41.05)

Figure 2.1.1: Plots of MOS vs HDR-VQM scores before and after INLSA alignment. The INLSA algorithm scales MOS values so that images which have similar objective scores also have similar MOS values. In order to compare the scatter plot quantitatively, the root mean squared error (RMSE) of the data is reported for each case.

other words, due to the characteristics of the experiments and test material, a similar value of the objective metric may correspond to a very different level of impairment in the subjective scale. Therefore, in order to use in a consistent way the MOS values of different subjective databases, these need to be mapped onto a common quality scale.

In order to align the MOS values of all five HDR image databases, we use the *iterated nested least square* algorithm (INLSA) proposed in [146]. This algorithm requires objective parameters for the alignment, under the assumption that those are sufficiently well correlated and linear with respect to MOS. Therefore, we selected the five most linear and most correlated objective quality metrics: HDR-VDP-2.2, HDR-VQM, PU-IFC, PU-UQI, and PU-VIF (more details about these metrics will be provided in the following). The INLSA algorithm first normalizes MOS scores from each source in the $[0,1]$ interval, and then aligns them by solving two least square problems: first, the MOS values of different datasets are corrected by an affine transformation in order to span the same subjective scale; second, the MOS values are aligned to the corresponding objective values by finding the optimal (in least-square sense) combination of weights such that the corrected MOS's can be predicted as a linear combination of objective parameters. These two steps, prediction and correction, are repeated iteratively till some convergence criterion is met. Details about the algorithm can be found in [146].

The scatter plots of MOS values and HDR-VQM metric values after alignment can be seen in Fig. 2.1.1(b). It can be observed that data points having similar HDR-VQM values have similar MOS values after INLSA alignment. After the alignment, all the MOS values have been mapped onto a common subjective scale, and they can be used in the evaluation of the objective quality metrics.

From Fig. 2.1.1(b), we notice that images in Database #2 [130] have very different characteristics compared to others, and MOS values are much more scattered than other databases after the alignment. This is mainly due to the characteristics of this database, i.e., the stimuli were mainly obtained by changing the tone mapping algorithm used in the compression, including many TMO's which produce strong color artifacts in the reconstructed HDR image. Also, different kinds of distortion are present simultaneously, such as color banding, saturation etc. In

some cases, it is noticed that false contours have been generated, and some color channels were saturated. Database #2 is thus very challenging for all the quality metrics we considered in this work. Therefore, in order to provide a complete overview of the performance of HDR fidelity metrics, in the following we report results both with and without including Database #2 in the evaluations.

After the alignment of MOS values of the databases, we obtain an image data set consisting of 690 (or 480 images if Database #2 is excluded) images compressed using JPEG, JPEG-XT, and JPEG 2000.

Considered objective metrics

In order to calculate quality metrics, we first scale pixel values to the range of luminance emitted by the HDR displays used in each subjective experiments. This is especially important for those metrics such as HDR-VDP 2.2 which rely on physical luminance. In order to compute these values, we convert HDR pixels into luminance emitted by a hypothetical HDR display, assuming it has a linear response between the minimum and maximum luminance of the display. As the same display (i.e. SIM2 HDR47E S 4K) has been used in all the experiments, we have selected the same parameters for all experiments, i.e., 0.03 cd/m^2 and 4250 cd/m^2 for minimum an maximum luminance, respectively.

We consider the following objective quality metrics:

- **HDR-specific metrics:** HDR-VDP-2.2 [131] and HDR-VQM [129] are fidelity metrics developed for HDR image and video, respectively. They model several phenomena that characterize the perception of HDR content, and thus requires some knowledge of viewing conditions (such as distance from the display, ambient luminance, etc.). The mPSNR is PSNR applied on an exposure bracket extracted from the HDR image, and then averaged across exposures.
- **Color difference metrics:** we use CIE ΔE 2000 (denoted as CIE ΔE_{00}), which entails a color space conversion in order to get perceptually uniform color differences [105], and its spatial extension [213] (denoted as CIE ΔE^S_{00}).
- **LDR metrics applied after a transfer function:** LDR metrics such as MSE, PSNR, VIF [172], SSIM [195], MSSIM [198], IFC [173], and UQI [197]. To compute these LDR metrics we use:
  - Physical luminance of the scene directly, denoted as *Photometric-*,
  - Perceptually uniform [5] encoded pixel values, denoted as *PU-*,
  - Logarithmic coded pixel values, denoted as *Log-*, or
  - Perceptually quantized [120] pixel values. For this case, only tPSNR-YUV has been considered as in [187].

Table 2.1.2: Pearson Correlation Coefficient (PCC) Results for Each Database and for Aligned Data

Metric	**Database #1**	**Database #2**	**Database #3**	**Database #4&5**	**Combined**	**Except Datab. #2**
Photometric-MSE	0.4051	0.1444	0.7080	0.5095	0.3651	0.6987
Photometric-PSNR	0.4409	0.2564	0.7132	0.5594	0.5166	0.6506
Photometric-SSIM	0.5016	0.3583	0.8655	0.6708	0.6441	0.7462
Photometric-IFC	0.7781	0.8234	0.9183	0.8195	0.8344	0.7680
Photometric-UQI	0.7718	0.8208	0.8846	0.7876	0.8312	0.7667
Photometric-VIF	0.7603	0.5076	0.8666	0.6144	0.6264	0.8452
PU-MSE	0.4824	0.3309	0.8559	0.8024	0.6273	0.7710
PU-PSNR	0.5297	0.3269	0.8606	0.8009	0.6271	0.7761
PU-SSIM	0.8661	0.7049	0.9532	0.9201	0.8441	0.9016
PU-IFC	0.7910	0.8422	0.9201	0.8566	0.8569	0.8024
PU-MSSIM	0.8847	0.7236	0.9564	0.9038	0.8570	0.9210
PU-UQI	0.7823	**0.8507**	0.8768	0.7777	0.8367	0.7637
PU-VIF	0.7845	0.7583	0.9349	0.9181	0.8574	0.8655
Log-MSE	0.6114	0.5314	0.8856	0.8820	0.6844	0.7872
Log-PSNR	0.6456	0.5624	0.8870	0.8819	0.7001	0.7923
Log-SSIM	0.8965	0.8035	0.9235	0.8255	0.8418	0.8401
Log-IFC	0.7919	0.8366	0.9167	0.8551	0.8530	0.8034
Log-UQI	0.7837	0.8268	0.8786	0.7830	0.8285	0.7592
Log-VIF	0.5079	0.6202	0.8354	0.7065	0.6049	0.6889
HDR-VDP-2.2 Q	**0.8989**	0.5482	0.9531	**0.9408**	0.7590	0.9261
HDR-VQM	0.8949	0.7932	**0.9612**	0.9332	**0.8807**	**0.9419**
mPSNR	0.6545	0.6564	0.8593	0.8587	0.7434	0.7959
tPSNR-YUV	0.5784	0.4524	0.8319	0.7789	0.6580	0.7718
CIE ΔE_{00}	0.6088	0.2553	0.7889	0.6082	0.4979	0.7752
CIE ΔE^{S}_{00}	0.6167	0.3331	0.8793	0.7322	0.5783	0.7929

2.1.2 Statistical analysis

We initially perform a statistical evaluation of the metrics listed above in terms of *prediction accuracy*, *prediction monotonicity*, and *prediction consistency*, following the recommendation [61]. For prediction accuracy, Pearson correlation coefficient (PCC), and root mean square error (RMSE) are computed. Spearman rank-order correlation coefficient (SROCC) is used to find the prediction monotonicity, and outlier ratio (OR) is calculated to determine the prediction consistency. These performance metrics have been computed after a non-linear regression performed on objective quality metric results using a logistic function, as described in the final report of VQEG FR Phase I [163].

For the sake of conciseness, we report only the results for PCC in Table 2.1.2 — the complete set of results is available in the original paper [210]. the results for the other performance criteria show essentially similar conclusions:

- The performance of many fidelity metrics may significantly vary from one database to another, due to the different characteristics of the test material and of the subjective eval-

uation procedure. In particular, Database #2 is the most challenging for all the considered metrics, due to its more complex distortion features, as discussed before.

- Despite the variations across databases, we can observe a consistent behavior for some metrics. Photometric-MSE is the worst correlated one, for all databases. This is expected as mean square error is computed on photometric values, without any consideration of visual perception phenomena. On the other hand, HDR-VQM, HDR-VDP-2.2 Q, and PU-MSSIM are the best performing metrics, with the exception of Database #2.
- In general, metrics which are based on MSE and PSNR (PU-MSE, Log-MSE, PU-PSNR, mPSNR, etc.) yield worse results compared to other metrics. Instead, more advanced LDR metrics such as IFC, UQI, SSIM, and MSSIM yield much better results. We also notice that mPSNR, tPSNR-YUV, and CIE ΔE 2000, which have been recently used in MPEG standardization activities, perform rather poorly in comparison to the others.
- Interestingly, the HDR-VQM metric, which has been designed to predict *video* fidelity, gives excellent results also in the case of static images, and is indeed more accurate on Database #2 than HDR-VDP-2.2.
- The impact of compression artifacts on image quality seems to dominate that of color differences, and CIE ΔE_{00} and CIE ΔE^{S}_{00} do not correlate well with subjective scores. Thus, our analysis confirms the results we have obtained in our work [206].

2.1.3 Discriminability analysis

The performance scores considered in the previous section assume that MOS values are known *deterministically*. Hence, the goal of fidelity metrics is to predict them as precisely as possible. However, MOS values are estimated from a sample of human observers, i.e., they represent expected values of random variables (the perceived annoyance or quality). Therefore, MOS are as well *random variables* which are known with some uncertainty, which is typically represented by their confidence intervals [56]. Thus, it can happen that images having the same MOS might have different underlying quality, and viceversa, the quality of images having different MOS could not be discriminated in practice. Therefore, in this section we propose another evaluation approach, which aims at assessing if an objective fidelity metric is able to discriminate whether two images have significantly different subjective quality.

Prior work has considered the problem of evaluating objective quality metrics taking into account the variability of MOS. Brill et al. [15] introduced the concept of *resolving power* of an objective metric, which indicates the minimum difference in the output of a quality prediction algorithm such that at least $p\%$ of viewers (where generally $p = 95\%$) would observe a difference of quality between two images. This approach has also been standardized in ITU Recommendation J.149 [60]. Nevertheless, this technique has a number of disadvantages that makes it little used in practice, e.g.: it requires transforming MOS to a common scale through a fitting that might be ill-posed; the resolving power might correspond to a variable metric resolution in the original scale; and the performance indicators are mainly qualitative and difficult to interpret. In parallel to our work, Krasula et al. [88] have proposed a different method which does not require using a common scale and uses instead a classification approach.

Our approach also does not need the transformation into a common scale, and overcomes the limitations of [15]. The basic idea of the proposed method is to convert the classical *regression* problem of accurately predicting MOS values, into a *binary classification* (detection) problem. We denote by $S(I)$ and $O(I)$ the subjective (MOS) and objective quality of stimulus I, respectively, for a certain objective quality metric. Given two stimuli I_i, I_j, we model the detection problem as one of choosing between the two hypotheses $\mathcal{H}_0$, i.e., there is no significant difference between the visual quality of I_i and I_j, and $\mathcal{H}_1$, i.e., I_i and I_j have significantly different visual quality. Formally:

$$\begin{aligned} &\mathcal{H}_0 : S(I_i) \cong S(I_j); \\ &\mathcal{H}_1 : S(I_j) \ncong S(I_j), \end{aligned} \tag{2.1.1}$$

where we use $\cong$ (resp. $\ncong$) to indicate that the means of two populations of subjective scores (i.e., two MOS values) are the same (resp. different). Given a dataset of subjective scores, it is possible to apply a pairwise statistical test (e.g., a two-way *t-test* or *z-test*) to determine whether two MOS's are the same, at a given significance level. In our work, we employ a one-way analysis of variance (ANOVA), with Tukey's honestly significant difference criterion to account for the multiple comparison bias.

In order to decide between $\mathcal{H}_0$ and $\mathcal{H}_1$, similar to Krasula et al. [88], we consider the simple test statistic $\Delta_{ij}^O = |O(I_i) - O(I_j)|$, i.e., we compare the difference between the objective scores for the two stimuli to a threshold τ, that is:

$$\text{Decide:} \begin{cases} \mathcal{H}_0 & \text{if } \Delta_{ij}^O \leq \tau \\ \mathcal{H}_1 & \text{otherwise.} \end{cases} \tag{2.1.2}$$

For a given value of τ, we can then label the set of stimuli as being equivalent or not. Clearly, the performance of the detector in (2.1.2) depends on the choice of τ. We call *true positive rate (TPR)* the ratio of images with different MOS's correctly classified as being of different quality, and *false positive rate (FPR)* the ratio of images with equal MOS's incorrectly classified as being of the different quality. By varying the value of τ, we can trace a Receiver Operating Characteristic (ROC) curve, which represents the TPR at a given value of FPR [80]. The area under the ROC curve (AUC) is higher when the overlap between the marginal distributions of Δ_{ij}^O under each hypothesis, that is, $p(\Delta_{ij}^O; \mathcal{H}_0)$ and $p(\Delta_{ij}^O; \mathcal{H}_1)$, is smaller. Therefore, the AUC is a measure of the *discrimination power* of an objective quality metric.

Table 2.1.3 reports the AUC results. In addition to the area under the ROC curve, we also compute the balanced classification accuracy:

$$Acc = \frac{2 \times TP}{TP + FN} + \frac{2 \times TN}{TN + FP}. \tag{2.1.3}$$

In Table 2.1.3 we report the maximum classification accuracy, $Acc^* = \max_\tau Acc$, which characterizes the global detection performance. These results in Table 2.1.3 are complemented with the percentage of correct decisions (CD) in [15], which is to be compared with Acc^*. We notice that, in general, the values of CD are much lower than Acc^*. This is due to the fact that the method in [15] not only aims at distinguishing whether two images have the same quality,

Table 2.1.3: Results of discriminability analysis: area under the ROC curve (AUC) and maximum classification accuracy. We report for comparison the fraction of Correct Decisions (CD) at 95% confidence level as proposed in [15]. For CD, '–' indicates that the 95% confidence level cannot be achieved.

	Combined			**Except Database #2**		
Metric	AUC	Acc*	CD [15]	AUC	Acc*	CD [15]
Photometric-MSE	0.532	0.530	–	0.644	0.614	0.317
Photometric-PSNR	0.576	0.556	–	0.633	0.596	0.249
Photometric-SSIM	0.609	0.590	–	0.677	0.633	0.306
Photometric-IFC	0.716	0.666	0.398	0.675	0.629	0.340
Photometric-UQI	0.765	0.707	0.380	0.730	0.678	0.296
Photometric-VIF	0.605	0.585	0.204	0.717	0.654	0.446
PU-MSE	0.596	0.580	–	0.677	0.645	0.379
PU-PSNR	0.625	0.593	–	0.715	0.661	0.380
PU-SSIM	0.721	0.663	0.399	0.804	0.725	0.512
PU-IFC	0.729	0.676	0.451	0.694	0.643	0.386
PU-MSSIM	0.737	0.680	0.434	0.838	0.758	0.598
PU-UQI	0.770	0.711	0.391	0.730	0.678	0.286
PU-VIF	0.782	0.719	0.463	0.802	0.735	0.493
Log-MSE	0.600	0.587	0.253	0.687	0.653	0.393
Log-PSNR	0.668	0.624	0.256	0.729	0.668	0.395
Log-SSIM	0.717	0.664	0.394	0.762	0.696	0.407
Log-IFC	0.725	0.673	0.443	0.694	0.642	0.382
Log-UQI	0.769	0.711	0.368	0.728	0.676	0.272
Log-VIF	0.634	0.593	0.217	0.666	0.635	0.282
HDR-VDP-2.2 Q	0.689	0.630	0.300	0.850	0.759	0.622
HDR-VQM	**0.791**	**0.727**	**0.487**	**0.893**	**0.816**	**0.684**
mPSNR	0.690	0.648	0.278	0.727	0.671	0.381
tPSNR-YUV	0.636	0.603	0.178	0.708	0.658	0.367
CIE ΔE_{00}	0.580	0.559	0.168	0.721	0.669	0.332
CIE ΔE^{S}_{00}	0.602	0.575	0.187	0.723	0.668	0.349

but also to determine which is the one with better quality. Thus the classification task is more difficult, as there are three classes – equivalent, better or worse – to label. Notice that in some cases, the CD cannot be computed, as the percentage of observers seeing a difference between image qualities is lower than 95% for any metric difference values.

The results of discriminability analysis lead to similar conclusions as the statistical analysis in Section 2.1.2. Nevertheless, even for the best performing metrics having correlation with MOS larger than 0.9, maximum accuracy saturates at 0.8. This suggests that there is still space for improving existing HDR objective quality measures, as far as discriminability (and not only prediction accuracy) is included in the evaluation of performance.

2.2 Blind HDR quality estimation disentangling perceptual and noisy features

Differently from the previous section, where we discussed fidelity (full-reference) quality metrics, in this section we take a *no-reference* perspective, i.e., we assume that the pristine image is not available. We continue to focus on assessing the distortion produced by HDR image compression algorithms.

We propose a model capable of predicting the perceived HDR image quality and localizing the distortions. We use a convolutional neural network (CNN) based architecture to achieve this goal. We approach the problem of designing a perceptual HDR no-reference image quality assessment (NR-IQA) model by dividing the visual quality analytic process into sub-components. We represent visual quality perception as the result of *two functional units*. The first unit takes a distorted image and detects error, and the second unit performs a perceptual scaling of this error to compute a quality score. By using a supervised learning approach, the mathematical behavior of these two units can be modeled. The data required for this training is obtained from an IQA dataset, which contains images and the corresponding quality scores.

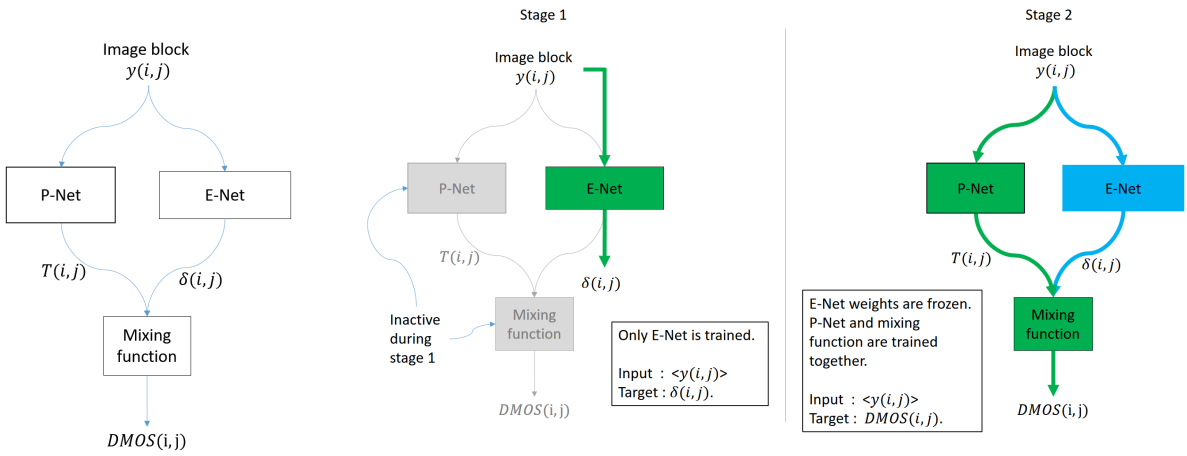
Specifically, the contributions presented include:

- Proposing an NR-IQA model based on a convolutional neural network architecture, which can separate pixelwise errors from their impact on perception in a distorted image. Our model outperforms other NR-IQA models and is competitive with state-of-the-art HDR full-reference IQA algorithms.
- Providing an accurate error prediction in a distorted image without a reference image.
- Predicting the visual masking effects without the need of explicit psychovisual subjective tests.

The content of this section is described in further details in the original paper [84].

2.2.1 Proposed model

To estimate HDR image quality, we design a system based on two convolutional neural networks (CNN), illustrated in Figure 2.1(a). The input are HDR image blocks composed of linear luminance values. We use a block size of 32×32 pixels. This is the same block size that was suggested in [69]. Our CNN model has three major parts: E-net, P-net and a Mixing function. E-net estimates the *Error* $\delta(i, j)$ of an image block centered at (i, j). P-net computes the *Perceptual Resistance* $T(i, j)$ of the block. The output of these two systems are then input to a *Mixing function*, to produce the local block quality. We obtain Differential Mean Opinion Scores (DMOS) for each image block. The block scores are then combined to generate the final image quality score. In our model, DMOS is a number directly proportional to the level of distortion in a HDR image. We describe each component in detail in the following.



(a) Proposed NR-IQA (b) Two-step training process

Figure 2.2.1: Proposed strategy for HDR NR-IQA. (a) E-net detects the error, P-net detects the perceptual resistance, and Mixing function consolidates the results and computes a DMOS. (b) The two-step training process enables to avoid the need for psychovisual ground-truth labels.

E-net Error Estimation

The Error $\delta(i,j)$ quantifies the change in statistics in a distorted image block. For an image block centered at (i,j), we define the error as,

$$\delta(i,j) = mean(\left|Y_R(i,j) - Y_D(i,j)\right|) \tag{2.2.1}$$

where Y_R and Y_D are, respectively, the original and distorted linear HDR luminance values of the image block centered at (i,j). This does not indicate a Full Reference computation, as the original version is only used during training (a pair of HDR image and its distorted version is used here). The objective is to train E-net with the distortion characteristics, like blocky artifacts, blurring effects, jagged edges, etc.

We use ℓ_1 norm for error computation (Eq. 2.2.1) instead of alternative measures such as ℓ_2 norm, to avoid over-emphasizing large errors. This is particularly important in HDR images where the histogram of Y is generally very skewed and some pixels take very high luminance values. We use our own CNN approach to design E-net to obtain and estimation $\hat{\delta}(i,j)$ of the error in Eq. (2.2.1)

P-net Perceptual Resistance

For each image block centered at (i,j), we compute the *Perceptual Resistance* $T(i,j)$. This value represents the difficulty for a viewer to perceive the error $\delta(i,j)$ of the block. A high T value implies that it is less likely to see the error, and hence the quality of the block is less affected (high perceptual resistance). Conversely, a low value implies that the image block will be perceptually degraded by error. Perceptual Resistance $T(i,j)$ aims to represent a combination

of all perceptual effects exhibited by an image block centered at (i, j). Though it is functionally similar to the pixel-wise just noticeable error measure used in conventional IQA systems like [215], [25] and [131], we introduce Perceptual Resistance as a new term because our model generates local quality scores (DMOS), as opposed to a local probability of error detection.

Instead of following the traditional perceptual modeling method of deriving perceptual thresholds from psychophysical experiments, we solve this problem by a data driven method. We use a convolutional neural network (CNN) based architecture, P-net, to derive the Perceptual Resistance of a block.

Mixing function

We use a *Mixing function* $f(\hat{\delta}, T)$, which combines the estimated error and Perceptual Resistance to generate a quality score. This is a critical part of the system because it is this value that is optimized by the training process to match human quality scores. The output of P-net would change based on how the Mixing function is designed.

The Mixing function is designed as follows, with error expressed in multiples of Perceptual Resistance:

$$DMOS = f(\hat{\delta}, T) = G\left(\frac{\hat{\delta}}{T}\right), \tag{2.2.2}$$

where G is a monotonically increasing function. By using this, we express error in *JND like* measure ($\frac{error}{JND}$), so that the error quantity is mapped onto a more perceptually relevant scale. Such interpretation is common in IQA literature, e.g., [215]

Mapping $\frac{\hat{\delta}}{T}$ to quality scores is achieved by the function G. Since increase in visible error always leads to decrease in quality and increase in DMOS, the latter must monotonically increase with $\frac{error}{JND}$, implying that G also has to be monotonically increasing with $\frac{\hat{\delta}}{T}$. Thus, any monotonically increasing function is sufficient for G. However, choosing a G that is too complex can lead to optimization problems because of unstable data points along the function, or low values for gradients, leading to slow or zero learning.

Based on the above considerations, we use $G(x) = 1 - \exp(-|kx|)$ and DMOS is computed as:

$$DMOS(i, j) = 1 - \exp\left(-\left|\frac{k * \hat{\delta}(i, j)}{T(i, j)}\right|\right) \tag{2.2.3}$$

This choice is inspired by the error model proposed in [215], but we introduce a scaling factor k. Here the added parameter (weight) k can be tuned during the training process, so that the predicted values of DMOS are as close to the ground truth DMOS as possible.

Network architecture

E-net is a typical CNN architecture consisting of five layers. Details are given in [84]. For P-net, we define a customized CNN layer, called augmented input layer. In this layer, in addition to

the original luminance values of the block, we compute the mean, variance and MSCN images. For the MSCN image, we use the formulation in [122], i.e., $MSCN(y_N(i,j)) = \frac{y_N(i,j)-\mu_{y_N(i,j)}}{\sigma_{y_N(i,j)}+0.01}$, where $\mu_{y_N(i,j)}$ is the mean and $\sigma_{y_N(i,j)}$ is the variance. They are computed by replacing every pixel (i,j) with the mean and variance, respectively, over a local Gaussian window of size N around (i,j).

2.2.2 Two-step training procedure

An important element in a CNN-based system is the selection of right labels for training. To force the desired behavior of the sub-components, we need to provide the right examples to each of the CNN's.

E-net detects blockwise errors. It is trained with linear luminance values of the distorted image as input, and per pixel errors (Eq. (2.2.1)) as output, which are available in the training stage.

For P-Net, the ideal training data is a numeric quantity, encapsulating all perceptual effects on the HVS, generated from an image block. Although we cannot get such a final value directly, our system can produce a quality score after the Mixing function process. We use this score for training. This two-stage training forces the P-net to extract a set of perceptual features from the image blocks and to derive a single final Perceptual Resistance value.

We therefore define our two-stage training process as follows:

Stage 1: E-net is trained with distorted image blocks as input and errors δ as target. The error is computed with Eq. (2.2.1).

Stage 2: all the training weights of E-net are frozen by setting their learning rate to zero. The whole network is then trained with image blocks as input and ground truth image quality of the whole image, $DMOS_{gt}$, as target. We use J as the cost function for any image block centered at (i,j), where

$$J(i,j) = \left|DMOS(i,j) - DMOS_{gt}\right|. \tag{2.2.4}$$

$DMOS(i,j)$ is the output of the Mixing function. The P-net and the mixing function (optimal value of k) get trained during this stage.

The overall process is illustrated in Fig 2.1(b).

Notice that in Eq. (2.2.4) we assume that the local quality of an image block is the same as the global image quality score, similarly to the setting in [69]. While this assumption is somehow inaccurate (as distortion can be unevenly spread across a picture), it has been proven to be accurate enough to predict image quality without reference [69].

2.2.3 Experimental results

We compare the performance of our algorithm with existing methods and show a clear improvement in performance. We conduct two tests: 1) a test of overall performance of the proposed

Table 2.2.1: Comparison of overall prediction performance. Highlighted in bold are the highest performing metrics.

(a) All data together

Scheme	Processing	SRCC	PLCC	RMSE
BRISQUE	Lin	0.7274	0.7231	18.1797
	PU	0.8047	0.7825	17.3576
	TMO - Drago	0.7374	0.7203	19.1261
	TMO - Reinhard 02	0.7782	0.7699	18.1523
	TMO - Reinhard 05	0.6903	0.6643	20.3307
	TMO - Mantiuk	0.6172	0.6148	22.1868
SSEQ	Lin	0.6022	0.6008	23.3017
	PU	0.7342	0.7175	19.4117
	TMO - Drago	0.6853	0.6954	20.8766
	TMO - Reinhard 02	0.6866	0.6688	21.0673
	TMO - Reinhard 05	0.6568	0.6467	20.5737
	TMO - Mantiuk	0.4185	0.4651	25.7570
BIQI	Lin	0.1817	0.1466	38.7513
	PU	0.3387	0.3445	30.5220
	TMO - Drago	0.2803	0.2960	41.0579
	TMO - Reinhard 02	0.3756	0.3766	33.2005
	TMO - Reinhard 05	0.3097	0.2874	27.7294
	TMO - Mantiuk	0.2822	0.2408	39.0999
DIIVINE	Lin	0.6677	0.6759	21.8020
	PU	0.7156	0.7193	18.7586
	TMO - Drago	0.7418	0.7400	18.9959
	TMO - Reinhard 02	0.7149	0.7024	20.7177
	TMO - Reinhard 05	0.7900	0.7809	17.2134
	TMO - Mantiuk	0.4946	0.4936	27.4918
kCNN	Lin	0.8363	0.8134	19.1753
	PU	**0.8638**	**0.8497**	**16.8937**
	TMO - Drago	0.7700	0.7485	18.2759
	TMO - Mantiuk	0.8075	0.8053	17.7948
	TMO - Reinhard 02	0.8613	0.8179	17.7157
	TMO - Reinhard 05	0.6438	0.6074	22.3484
Proposed	PU	**0.8860**	**0.8871**	**16.4171**
Proposed	Lin	**0.8920**	**0.8860**	**14.1464**

(b) Cross-dataset

Scheme	Processing	SRCC	PLCC	RMSE
BRISQUE	Lin	0.5400	0.4772	28.8475
	PU	0.7135	0.6503	20.5534
	TMO - Drago	0.6337	0.5903	21.7118
	TMO - Reinhard 02	0.6583	0.6512	18.4500
	TMO - Reinhard 05	0.3524	0.3946	30.6615
	TMO - Mantiuk	0.5887	0.5493	22.7529
SSEQ	Lin	0.5287	0.4714	25.2588
	PU	0.6492	0.6111	19.6977
	TMO - Drago	0.5865	0.5634	22.6987
	TMO - Reinhard 02	0.5810	0.5644	22.9900
	TMO - Reinhard 05	0.4990	0.5036	24.9193
	TMO - Mantiuk	0.4973	0.4770	21.2044
BIQI	Lin	0.2845	0.2831	31.0686
	PU	0.4386	0.4399	21.2084
	TMO - Drago	0.5332	0.4436	25.6200
	TMO - Reinhard 02	0.4632	0.4358	22.0376
	TMO - Reinhard 05	0.5748	0.5630	19.4825
	TMO - Mantiuk	0.4651	0.4571	24.2268
DIIVINE	Lin	0.5041	0.5209	20.6506
	PU	0.5318	0.5442	19.6772
	TMO - Drago	0.4143	0.4065	25.9697
	TMO - Reinhard 02	0.3634	0.3953	26.1464
	TMO - Reinhard 05	0.5558	0.5374	19.3122
	TMO - Mantiuk	0.4138	0.4496	21.0499
kCNN	Lin	0.6991	0.7008	19.3677
kCNN	PU	0.7694	0.7544	18.5854
Proposed	Lin	**0.8672**	**0.8780**	**18.626**
HDR-VDP	Full Reference	**0.9298**	**0.9408**	**10.120**
HDR-VQM	Full Reference	**0.9193**	**0.9332**	**10.725**
PU-MSSIM	Full Reference	**0.8969**	**0.9038**	**12.775**
PU-SSIM	Full Reference	**0.9121**	**0.9201**	**11.688**

method on a large dataset of subjectively annotated HDR images; 2) a cross-dataset test to assess the generalization capabilities of the proposed approach.

We use the HDR dataset described in Section 2.1, obtained by aligning the five subjectively annotated HDR image datasets as shown in Table 2.1.1 and in Figure 2.1.1. The datasets provide only MOS values of the images. Since our system requires the difference of mean opinion scores (DMOS), we convert MOS to DMOS as follows:

$$DMOS_{gt}(i) = \frac{MOS_{MAX} - MOS(i)}{MOS_{MAX}}, \tag{2.2.5}$$

where $DMOS_{gt}(i)$ is the ground truth DMOS score for image i, MOS_{MAX} represents the maximum MOS in the IQA training dataset and $MOS(i)$ is the MOS of the ith image of combined database after INLSA alignment.

We compared our approach with a number of state-of-the-art LDR NR-IQA methods: BRISQUE [122], SSEQ [102], BIQI [124], DIIVINE [125], and kCNN [68], with and without pre-processing

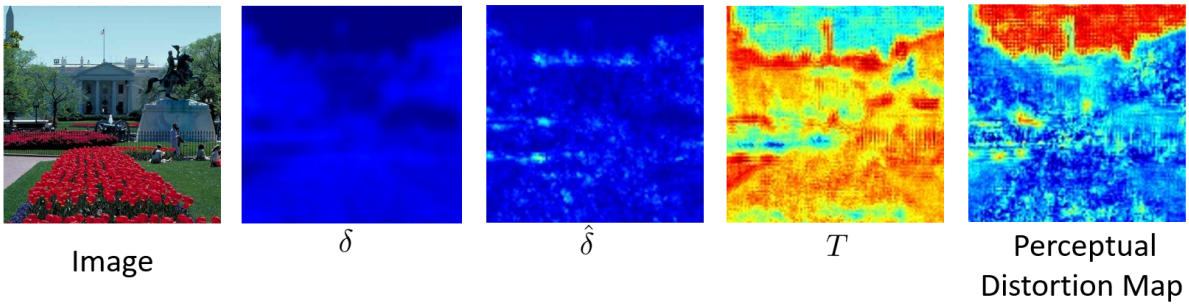


Figure 2.2.2: Example output of E-net and P-net. Given the input image, we report: the actual error δ, the output of E-net $\hat{\delta}$ (estimated error), the output of P-net T (perceptual resistance), and the estimated local perceptual quality $DMOS(i, j)$ after the mixing function..

operators. PU encoding [5] was used as a pre-processing operator. We also used a number of tone-mapping operators, which include [159], [157], [30] and [113], in pre-processing.

In order to test the overall performance of the proposed method, we (re)train each algorithm on our combined image dataset, by splitting it in training/testing subsets (80% for training and 20% for testing), in such a way that the training and testing sets do not contain the same contents. We repeat this procedure 100 times, and report median prediction performance in Table 2.1(a). The best performances are obtained by using BRISQUE [122] and kCNN [69]. The high performances of BRISQUE and kCNN on HDR linear values can be attributed to the features they use, i.e., the MSCN coefficients. It is likely that the normalization by variance cancels the effects of the increased dynamic range and yields a similar distortion pattern as LDR images.

Furthermore, we observe a clear performance improvement in LDR NR-IQA algorithms if the data is pre-processed and the dynamic range of the data is reduced to LDR levels, confirming what we found in Section 2.1. PU encoding improves the performance in most of the cases. The best performance among LDR NR-IQA is obtained while using PU encoding in conjunction with kCNN. The performance of the proposed system is significantly better than the other algorithms in all cases both with or without pre-processing using PU encoding.

In order to demonstrate the generalization capabilities of the proposed NR-IQA technique to different conditions and contents, we train the algorithms using datasets #1, #2 and #3, and test them on datasets #4 and #5 (see Table 2.1.1). Notice that this is a more challenging scenario, and in fact the performance of the metrics are generally lower. We report also the performance of some of the best full-reference metrics from Table 2.1.2 for comparison. The proposed algorithm outperforms related methods in all test cases when considering generalization to a real-world scenario, and achieves performance close to full-reference quality metrics.

Figure 2.2.2 reports a visual example of the output of the proposed network for a given compressed image. Notice that the perceptual resistance, T, is higher in areas which are either darker (lower luminance reduces the sensitivity to artefacts) or have strong textures, such as the flowers (contrast masking), while it is lower in smoother regions such as the sky, where distortion is more visible. More examples and analyses are reported in the paper [84].

We have shown in later work [85] that the proposed model with E-net and P-net, and the cor-

responding two-step training, is able to learn somehow contrast sensitivity from image quality, also for the general case of LDR images. Our experiments in [85] demonstrate that the latent information about distortion visibility carried by supra-threshold quality scores can be recovered and used to predict near-threshold local masking. One advantage of our approach, compared to models based on psychophysical data, is that it can leverage the larger availability of subjectively annotated image quality datasets.

2.3 Perceived dynamic range for HDR images

In this section we consider measuring a specific *perceptual attribute* of an HDR picture: its dynamic range. Specifically, we consider this problem from a *no-reference* perspective, where the perceptual attribute is an intrinsic property of the image and not one deduced by a difference to a reference picture with an "ideal" quantity of that attribute.

One of the main reasons why high dynamic range is supposed to boost the quality of visual experience is its ability to reproduce very bright and very dark portions of a scene concurrently. The span between these extrema in the brightness scale is commonly referred to as the *dynamic range* of a picture. The dynamic range of image or video content is typically computed as the ratio between the maximum and minimum pixel luminance of an image, which will be referred to as *pixel-based* dynamic range (DR) in the rest of this section. Such a computation can be biased due to image noise or singularities, such as isolated pixels with extreme luminance values. Furthermore, such measures do not capture the complex behavior of the human visual system's (HVS) response and perception of lightness [41]. Instead, the **perceived** dynamic range (PDR) depends on more complex characteristics of the content, and its assessment in HDR conditions is relevant in a number of applications, from optimization and assessment of inverse tone mapping operators (ITMOs) [26], to the evaluation of HDR displays and HDR content selection for subjective studies [135].

In this section we describe a study to model the PDR for HDR images. Specifically, our methodology consists in: i) collecting a dataset with PDR mean opinion scores for chromatic and achromatic HDR pictures; ii) proposing and selecting a number of features to explain PDR, and formulating a model with them. In summary, the contributions of this work are the following:

- we create a subjectively annotated data set with PDR values, using complex, chromatic and achromatic stimuli and HDR viewing conditions using an HDR display;
- we propose a novel test methodology for measuring perceived dynamic range, partially inspired by the subjective assessment methodology for video quality (SAMVIQ) [12];
- based on the results of the study, the Pearson's correlations between mean opinion scores (MOS) and five image features are analyzed;
- the effect of chromatic information on perceived dynamic range is investigated and the relation between chromatic and achromatic quantified;
- we propose a model for predicting the perceived dynamic range for both achromatic and chromatic.

The content of this section is described in greater detail in the original papers [52, 48].

2.3.1 Subjective dataset

Experimental design

In the study, a subjective evaluation of PDR of both achromatic and chromatic images was conducted. The participants were asked to evaluate the *overall impression of the difference between the brightest and the darkest part(s) in the images*. The independent variable was the image content, while the dependent variable was the reported PDR of the image. The study was conducted in two separate sessions, one for achromatic and another for chromatic images.

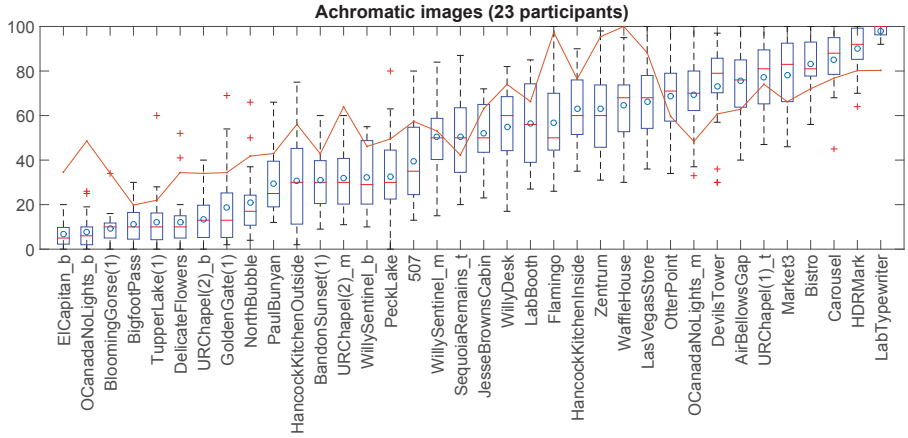
Three possible evaluation methods are typically considered during the design of experiments of this type: paired comparison, ranking and rating. Paired comparisons were ruled out due to their impracticality with large data sets. More efficient pair comparison techniques can be used under certain assumptions. However, due to multidimensionality and non-deterministic DR appearance, these assumptions in our case were violated. While the ranking methods are straightforward, and quick to conduct, as with pairwise comparisons, they provide no information on the magnitude of the differences. Therefore, this method has been designed in order to use the advantages of the three methods: it permits ranking of the stimuli, a direct comparison between the image pairs, and it uses the continuous scale for subjective scores.

The evaluation method was inspired by the Subjective Assessment Methodology for Video Quality (SAMVIQ) [12], adapted to static images. The data set consisted of 36 images, selected from the pool of 137 images, and divided into three subsets of 12 pseudo-randomly selected images in a randomized order. The evaluation session was not time constrained. Each subset was evaluated independently, allowing participants to re-evaluate any image within, but not across subsets as many times as they wanted. This allowed for multiple comparisons between the images and fine adjustments of the scores. The rating was performed on a 0-100 continuous scale, divided into five equal intervals with corresponding labels: very low, low, medium, high and very high, included for general guidance.

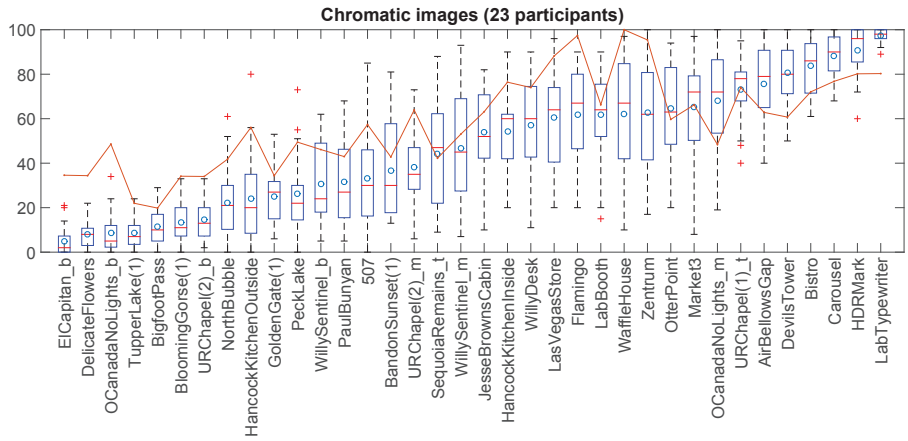
Stimuli

A set of 33 images were selected from the HDR Photographic Survey [33] in order to have an approximately uniform distribution of three image features: dynamic range, image key, and spatial information, as described in [52]. Since most of the images from the Fairchild's data set are photographs of nature, a single frame from the *Market* HDR video sequence proposed in MPEG by Technicolor [93] and a frame from both the *Carousel* and *Bistro* video sequences from the Stuttgart HDR Video Database [38] were added to the test data set. All 36 images were converted to the corresponding achromatic images, using BT.709 primaries to compute relative luminance [59].

The images were reproduced on an HDR SIM2 HDR47ES4MB 47" display, employed in the DVI Plus (DVI+) mode, that allows for directly and independently controlling backlight LEDs



(a) Achromatic



(b) Chromatic

Figure 2.3.1: Extended boxplot diagrams for achromatic (top) and chromatic (bottom) images. Blue circles = MOS; Red horizontal lines = median values; Blue boxes = the interquartile ranges; Whiskers = adjacent values; Red crosses = outliers; Red line: pixel-based DR values (scaled as $DR = \frac{DR}{\max(DR)} \cdot 100$). The scores are sorted by the mean value.

and LCD pixel values, based on the dual-modulation algorithm [209].

Results

In Figure 2.3.1, the extended box plot depicts the distribution of the perceived DR scores, with the corresponding mean and median values, confidence intervals and outliers. In addition, the pixel-based DR is also presented to visually display the correlation between the subjective and objective scores. The Pearson and Spearman correlations between the chromatic MOS and achromatic MOS are $r = .99$ and $r_s = .98$, respectively, indicating the chromatic and achromatic scores are very highly correlated.

2.3.2 Considered image features

In order to compute image features, the pixel luminance values were first scaled to the display range, yielding physical luminance in cd/m^2, with the following equation:

$$L' = \frac{L - \min(L)}{\max(L) - \min(L)} \cdot (Disp_{\max} - Disp_{\min}) + Disp_{\min}, \tag{2.3.1}$$

where $Disp_{\min} = 0.03\,cd/m^2$ and $Disp_{\max} = 4250\,cd/m^2$ in our setup.

We consider the following 5 features as factors that might explain the PDR:

- **Pixel-based Dynamic Range (DR)**, calculated after excluding 1% of the darkest and brightest pixels in the image, using

$$DR = \log_{10} \frac{\max(L')}{\min(L')}, \tag{2.3.2}$$

  where L' is the image with scaled values.

- **Image key (IK)**, which gives a measure of the average image brightness, defined as:

$$IK = \frac{\ln(\mathrm{avg}(L)) - \ln(\min(L))}{\ln(\max(L)) - \ln(\min(L))}, \tag{2.3.3}$$

  where the $\mathrm{avg}(L)$ was computed as $\ln(\mathrm{avg}(L)) = \Sigma_{ij} \ln(L(i,j) + \delta)/N$, with $\delta = 10^{-5}$ to avoid singularities and N was the number of pixels. Once again, $\min(L)$ and $\max(L)$ were calculated robustly, after excluding 1% of the darkest and the brightest pixels.

- **Area'** of specular highlights, calculated as $Area' = Area^{1/4}$, where:

$$Area = \sum_{ij} (L(i,j)), \text{ for } L(i,j) > 2400\,cd/m^2, \tag{2.3.4}$$

  is the number of pixels greater than the diffuse white threshold value. The value of 2,400 cd/m^2 was selected as recommended in ITU document [58], in accordance with the results of the study by Daly et al. [24], where different thresholds were found for different levels of user expertise. The $1/4$ power has been introduced a posteriori to linearize the relation of the feature with the collected MOS.

- **Contrast (C)**, which we compute using an adapted version of the multi-scale contrast measure of Peli [140] — details about computation are available in the original paper [48].

- **Colorfulness (Col)**, only for chromatic stimuli, computed as in [46].

The Pearson and Spearman correlation coefficients between each of these features and MOS values are reported in Table 2.3.1.

Table 2.3.1: Pearson's r and Spearman's r_s correlation coefficients between MOS values for both achromatic and chromatic images and five objective measures: DR, IK, Area, C and Col. * denotes significance at $p < .01$.

MOS	**Achromatic**				**Chromatic**				
Measure	DR	IK	Area'	C	DR	IK	Area'	C	Col
r	.87*	-.60*	.87*	-.19	.84*	-.61*	.87*	-.22	-.47*
r_s	.87*	-.55*	.89*	-.24	.84*	-.57*	.90*	-.27	-.43*

2.3.3 PDR predictor model

We use a multi-variate linear regression to produce a model that can predict the PDR based on the features described above. First, we normalize the features values to put them in the same scale:

$$x_i' = \frac{x_i - \frac{1}{n}\sum_{i=1}^{n} x_i}{\max(X) - \min(X)} \tag{2.3.5}$$

where x denotes a given feature. Then, we use a hierarchical approach in order to find the independent variables (image features) that significantly improve the prediction of the outcome variable (PDR). Since pixel-based DR is known to be a good predictor of the perceived DR, it is selected for the first block in the hierarchy. All other predictors (IK, Area', C and Col) are added to the second block, and the contribution of each predictor is computed by looking at the semi-partial correlation with the outcome. After this procedure, only the Area' is found to significantly contribute to PDR prediction. Therefore, only DR and Area' are retained to create the model.

Once the two relevant features are selected, the PDR for the achromatic images is predicted as:

$$\widehat{PDR} = 0.573 \cdot DR + 0.448 \cdot Area' \tag{2.3.6}$$

while for the chromatic images it is:

$$\widehat{PDR} = 0.506 \cdot DR + 0.471 \cdot Area'. \tag{2.3.7}$$

Notice that the two models are indeed very similar, which confirms the high correlation between chromatic and achromatic PDR.

From the point of view of model prediction performance, we observe that the correlation between predicted and ground-truth PDR is $r = .945$ and $r = .932$ for achromatic and chromatic images, respectively. If only DR was considered, the correlation would have been significantly lower, i.e., $r = .866$ and $r = .839$, respectively. An analysis on eight scenes with the highest discrepancies between the MOS and pixel-based DR values, described in more detail in [48], shows that the PDR prediction is significantly improved when the Area predictor contributes to the model.

Although the results show that, overall, the PDR prediction with the proposed model is closer to the MOS it is likely that there will be images where this is not the case due to the excessive complexity of the HVS and the related processes in the perception of such visual attributes. We

have shown in [50] that first-order statistics (i.e., the histogram) of an image are not sufficient to account for the perception of dynamic range. A limitation in this domain is the lack of annotated data, especially for HDR conditions. Another interesting direction to explore is modeling PDR for dynamic (video) content, where also more complex adaptation mechanism enter into play.

2.4 Towards a unified quality scale fusing rating and ranking measures

Differently from previous sections, where we focused on *objective* quality assessment, in this section we change perspective and we consider how to efficiently collect *subjective* ground-truth scores. This is especially important, since collecting subjective scores is an expensive and time-consuming task, and quantifying precisely perceptual experience enables to better employ a given budget of measurements and to merge existing datasets collected with different methodologies.

Two of the main methods of subjective quality assessment for multimedia content are direct rating and ranking. Direct rating methods ask the observers to assign scores to observed stimuli. They may involve displaying a single stimulus (absolute category rating (ACR), single stimulus continuous quality evaluation (SSCQE)), or displaying two stimuli (double stimulus impairment scale (DSIS), double stimulus continuous quality evaluation (DSCQE)). Ranking methods ask the observers to compare two or more stimuli and order them according to their quality. The most commonly employed ranking method is pairwise comparisons (PWC).

Essentially, rating has the advantage to provide an interpretable, supra-threshold scale of quality or distortion impairment, but it also requires a careful training of subjects, who might have a different interpretation of the scale adjectives. As a consequence, the rating scale is in general not universal. On the other hand, pairwise comparison experiments have a lower cognitive load, require little training and generally eliminate the bias of the observer. However, the total number of possible comparisons increase quadratically with the number of stimuli, which makes a full comparison approach unfeasible. In practice, not all comparisons are equally useful, e.g., comparing stimuli with too close or too distant impairment levels is generally uninformative [199]. Pairs of stimuli to be compared can be sampled iteratively based on the previously compared stimuli, based on heuristics [149] or, information-theoretic criteria[203].

Motivations and contributions of this work

The vast majority of studies employing the pairwise comparison method compare only the images depicting the same content, for example comparing different distortion levels applied to the same original image. However, assessing and scaling each content independently makes it difficult to obtain scores that correctly capture quality differences between conditions *across different contents* on a common quality scale. Furthermore, pairwise comparison capture only relative quality relations. Therefore, in order to assign an absolute value to such relative measurements, the experimenter needs to assume a fixed quality for a certain condition which is

then used as reference for the scaling. As a result, the scaling error accumulates as conditions get perceptually farther from the reference.

On the other hand, it is useful in practice to aggregate quality scores obtained from different quality evaluation experiments, e.g., to create larger annotated datasets. While this aggregation of subjective quality scores is usually done for rating (i.e. mean opinion scores) [146, 147, 210] or pairwise comparisons [145, 174] individually, little has been done to study the fusion of scores obtained by both these two methodologies.

In this section we consider the above two questions — how to obtain a unified quality scale merging rating and pairwise comparisons, and which is the role played by cross-content comparisons. Specifically, we describe the following contributions:

- We show the importance of scaling PWC results according to a given observer model, in order to convert preferences to an interval scale. Differently from typical mean opinion score (rating) scales, the obtained scale can be interpreted in terms of probability of preferences;
- We find experimentally that the relation between rating and scaled PWC results is approximately linear. This can be used to fuse measurements obtained from the two types of experiments;
- We show the advantages of adding cross-content comparisons in PWC, and we analyze through simulations which is the optimal ratio of cross/same content comparisons in a typical subjective quality assessment test.

The content of this section is presented and discussed in greater details (especially, with more experimental results and analyses) in the papers [207, 211, 142].

2.4.1 Observer model

In order to map data collected in experiments into a unified quality scale, we need to make certain assumptions about how observers respond. Such assumptions are encapsulated in the *observer model*. Observers might vary in their notions of quality among them (inter-observer variance), and their opinions are also likely to change when they repeat the same experiment (intra-observer variance). Thus, quality is not a deterministic value, but a random variable, which accounts for the subjective nature of these experiments. We describe in the following the observer models we employ for rating and pairwise comparison experiments.

Rating

In *rating* experiments the random variable associated with the quality can be expressed using the following model of observer rating behavior [64]:

$$\pi_{ik} = m_i + \delta_k + \xi_{ik}, \tag{2.4.1}$$

meaning that the rating π_{ik} for observer k and condition i depends on: m_i, the ground truth quality score; δ_k, the subject bias; and ξ_{ik} the subject inaccuracy and stimulus scoring difficulty. All components in the model are assumed to be independent random variables that are Normally distributed and ξ_{ik} is assumed to have a zero mean. This makes rating π_{ik} also Normally distributed.

Pairwise comparisons

As for *pairwise comparisons*, the two most widely used observer models are Thurstone [183] and Bradley-Terry [14]. In practice, both lead to similar solutions. Within the Thurstone model the perceived quality of condition i is modeled as a random variable:

$$\omega_i \sim N(q_i, \sigma_i) \tag{2.4.2}$$

where the mean of the distribution is assumed to be the (latent) true quality score q_i and the standard deviation σ_i accounts for combined inter- and intra-observer variance. Individual quality scores of compared conditions can be inferred from the relative distances, calculated as:

$$\omega_j - \omega_i \sim N(q_{ij}, \sigma_{ij}) \tag{2.4.3}$$

where σ_{ij} is the standard deviation of a new distribution obtained from the difference between two quality distributions and $q_{ij} = q_i - q_j$. A typical assumption, known as Thurstone Case V, assumes σ_{ij} to be the same across all conditions, and thus $\omega_i \sim N(q_i, \sigma)$.

The main difference between Thurstone Case V and Bradley-Terry models is that in the latter the difference between quality scores is expressed using a logistic distribution instead of a normal distribution. This leads to a more efficient numerical solution when optimizing quality scores. The difference between the cumulative Gaussian and logistic distribution is displayed in Figure 2.1(a), which shows that the two models are indeed very similar. In the following, we employ the Thurston Case V model, though the proposed approach might be extended to the Bradley-Terry one.

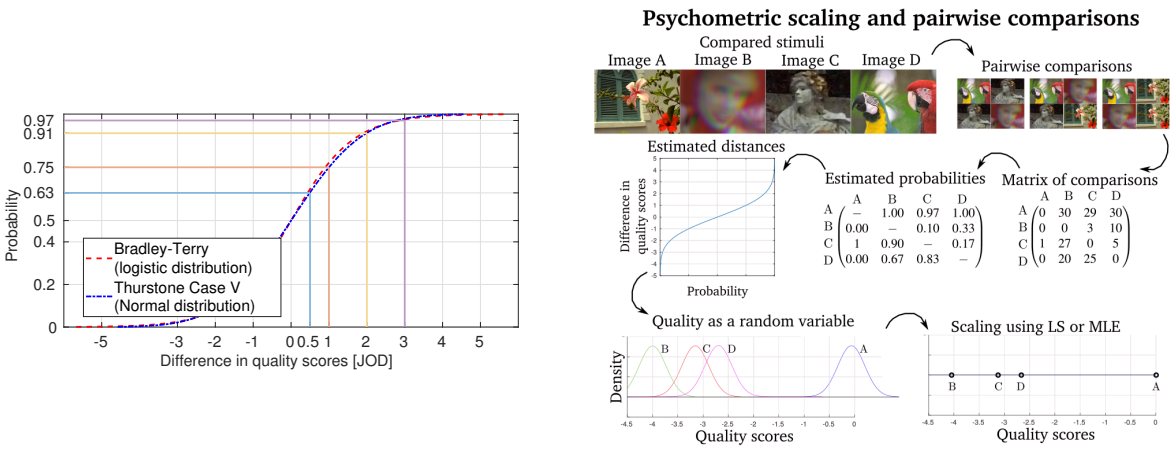
2.4.2 Psychometric scaling

Figure 2.1(b) shows a graphic representation of different steps in psychometric scaling via pairwise comparisons. Psychometric scaling aims to estimate the latent scores $\hat{\mathbf{q}}$ such that distances between scores closely resemble distances $\hat{q}_i - \hat{q}_j$.

To this end, the results of a pairwise comparison experiment are first arranged in a matrix $\mathbf{C}$, in which element c_{ij} counts the number of times stimulus i was chosen as better than j. Probabilities p_{ij} of $\omega_i > \omega_j$ can be empirically estimated:

$$\hat{p}_{ij} = \frac{c_{ij}}{c_{ij} + c_{ji}}, \ i \neq j. \tag{2.4.4}$$

Scaling consists in recovering the distance $q_i - q_j$ between underlying quality scores q_i and q_j,



(a) (b)

Figure 2.4.1: (a) Different cumulative distributions mapping probabilities into distances in the scale. Parameters for Thurstone and Bradley-Terry models were chosen such that the difference in 1 unit correspond to 75% probability of one condition being better than another. (b) Examples of different subjective judgment experiments and graphic representation of scaling using pairwise comparisons.

given an observer model. Following Thurstone Case V assumption, the difference of ω_i and ω_j is a Gaussian random variable (for Bradley-Terry, it is a logistic), as shown in Eq. (2.4.3). The probability of choosing condition i over j can be computed using the cumulative distribution over the difference $\omega_i - \omega_j$:

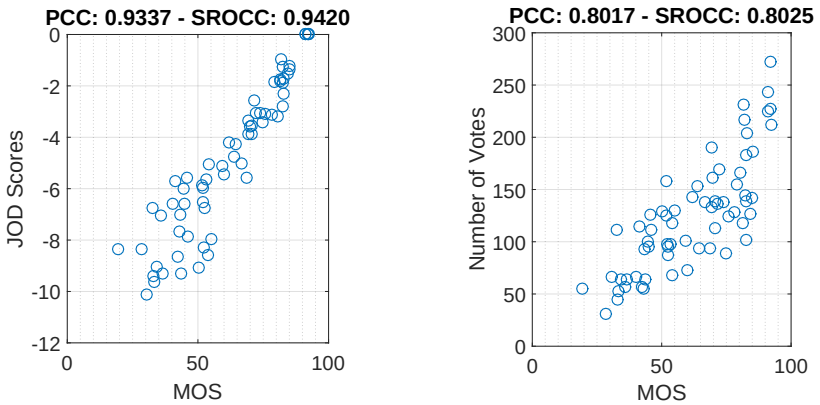
$$P(\omega_i > \omega_j) = F(q_{ij}, s_{ij}) \approx \hat{p}_{ij}, \tag{2.4.5}$$

where F is the cumulative distribution function associated to the chosen observer model and s_{ij} the parameter associated to the distribution (σ_{ij} for the Normal distribution in Thurstone model and s_{ij} for the logistic function in Bradley-Terry model). $P(\omega_i > \omega_j)$ is approximated using $\hat{p}_{ij}$. The inverse of F is shown in Figure 2.1(a). Note that the choice of s_{ij} determines the relationship between distances in the quality scale and probabilities of better perceived quality.

The probability of observing pairwise comparisons c_{ij} given latent quality scores q_i is explained by the Binomial distribution:

$$P(\mathbf{C}|\mathbf{q}, \sigma) = \prod_{i,j} \binom{n_{ij}}{c_{ij}} F\left(q_{ij}, s_{ij}\right)^{c_{ij}} \left(1 - F\left(q_{ij}, s_{ij}\right)\right)^{n_{ij} - c_{ij}}, \tag{2.4.6}$$

where $n_{ij} = c_{ij} + c_{ji}$ and F is the cumulative distribution from Eq. (2.4.5). Under Thurstone Case V assumptions, F is the cumulative normal distribution and $s_{ij} = \sqrt{2}\sigma$, where σ is the standard deviation of the observer model. σ is often selected so that when conditions are 1 unit apart in the quality scale, 75% of observers select one condition over another. This corresponds to $\sigma = 1.0484$ and $s_{ij} = 1.4826$ for normal distribution. We call this unit on the obtained scale a **Just Objectionable Difference (JOD)**, in contrast to the commonly used Just Noticeable Difference (JND), to emphasize the fact that here we consider quality preferences (stimulus



(a) JOD vs. MOS (b) VC vs. MOS

Figure 2.4.2: Psychometric scaling (a) vs. vote counts (b), on the HDR video dataset in [207]. In both cases, the PWC scores are compared to MOS obtained by a rating experiment on the same stimuli. The JOD scores are more in agreement with those of the rating experiment.

i is better than j), rather than detection of differences (i is different from j). Assuming that all the undistorted reference stimuli are equivalent to each other (i.e. having pristine quality with "0" quality score), the distorted images would then have negative JOD quality values after scaling, corresponding to the distortions compared to the undistorted reference stimuli (unless enhancement is considered).

Given the posterior probability in Eq. (2.4.6), the latent quality scores $\mathbf{q}$ can be found using the maximum likelihood estimation [141].

Scaling vs. vote counts

It should be noted that in some works the scaling of quality scores is avoided and the quality estimates are computed directly by summing up columns (or rows) of the comparison matrix. For example, the quality scores for the TID2013 dataset were computed as the average number of votes (wins in pairwise comparisons) that each condition received [149]. We refer to this approach as vote counts (VC). Such an approach works only if each condition was compared the same number of times and it is unsuitable for imbalanced experiment designs.

We verified experimentally in [207] that psychometric scaling, in general, yields subjective scores which are more coherent across experiments. Specifically, we simulated vote counts based on the comparison matrix obtained in a PWC experiment, and we correlated with the results of a rating experiment conducted on the same dataset. The scatter plots in Figure 2.4.2 shows that JOD values are more in agreement with mean opinion scores.

2.4.3 The relation between MOS and PWC and the importance of cross-content comparisons

Now that we have introduce psychometric scaling and shown the importance of the observer model in converting empirical probabilities to distances, we move to study the relationship between scaled PWC results (expressed in JOD) and rating results, expressed on a mean opinion score scale. To this end, we employ the HDR video quality database (HDRDVB) proposed in [206], which has been augmented with further rating and PWC measurements in [207]. Similar analyses as the one conducted here have been carried out on other image and video quality datasets in [142, 211], but are not reported here for the sake of conciseness.

HDRVB dataset

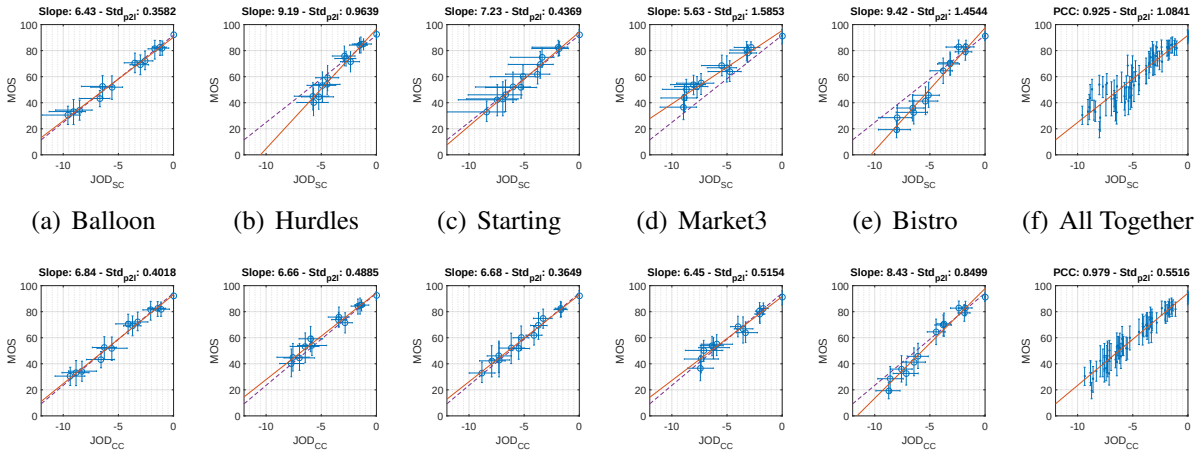
This dataset includes subjective scores for a total of 60 distorted stimuli from 5 original contents. Originally created to analyse the effect of different colourspaces on HDR compression, this database contains subjective quality scores collected using 4 different subjective experiment sessions and includes:

1. Double stimulus impairment scale (DSIS) session
2. PWC with only same-content pairs
3. Additional PWC with cross-content pairs
4. Additional PWC with same-content pairs

In total, the stimuli were compared 6390 times (5190 same-content and 1200 cross-content). The preference matrices of the PWC experiments were found and JOD scores were estimated using three different sets of PWC data. $JOD_{Standard}$ was found using the data acquired in the same-content PWC experiment. JOD_{CC}, on the other hand, was found using the data acquired in both the same-content and the cross-content PC experiments. Finally, JOD_{SC} is obtained by adding to the original PWC same-content data additional same-content comparisons, to have a total number of conditions comparable to those used for JOD_{CC}. For the DSIS experiment, the MOS values were calculated by taking the mean of opinion scores. Confidence intervals (CI), on the other hand, were calculated using bootstrapping in order to compare them to the CIs of JOD scores. These JOD scores are plotted vs. MOS values in Figure 2.4.3.

Linear relationship between MOS and JOD

The results in Figure 2.4.3 show that there is a strong relationship between MOS values and JOD scores. The introduction of cross-content pairs increases the correlation and linearity of the relationship between JOD and MOS. To be exact, the Pearson correlation coefficient (PCC) was found to be $\rho = 0.925$ for the case with only same-content pairs and $\rho = 0.979$ for the case including both same-content and cross-content pairs.



(a) Balloon (b) Hurdles (c) Starting (d) Market3 (e) Bistro (f) All Together

(g) Balloon (h) Hurdles (i) Starting (j) Market (k) Bistro (l) All Together

Figure 2.4.3: Relation between JOD and MOS and effect of cross-content comparisons. (a)-(f): JOD_{SC} vs. MOS. The solid red line indicates the best linear fit to the data, and the dashed violet line indicates the best linear fit line of the case 'All Together'. (g)-(l): JOD_{CC} vs. MOS. Instead of only same-content pairs, a combination of same-content and cross-content pairs were used to find JOD_{CC}. Figure best seen on a display.

Reduced Content Dependency

In Figure 2.4.3, the slopes of the best fitted line are found for each content. In order to find the effect of the addition of cross-content pairs, the variance of these slopes was found. Variance of the slopes in the case of JOD_{SC} was 2.7972 and in the case of JOD_{CC} was 0.6445. Another metric, Std_{p2l}, was computed for each figure presented. It is calculated as $Std_{p2l} = \sqrt{mean(d(P,l)^2)}$ where $d(\cdot)$ is the perpendicular distance from point P to line l. In the case of Figures 2.4.3(a)-(e) and (g)-(k), Std_{p2l} was computed considering the dashed violet line, i.e., the best linear fit when all the contents are considered together. It is clear that the addition of cross-content pairs decrease the variance of the slopes of the best fitted line for each content and Std_{p2l} as well, thus bringing JOD scores closer on a common quality scale.

Reduced Error Accumulation

In order to analyze the change in CI, average CI values are reported in Table 2.4.1. Since the CI does not change with respect to the color space much, the CI values were averaged for the same bitrate. The last column of Table 2.4.1 shows that the CIs are decreased for almost every case up to 30-60%, especially at higher bitrates where scaling error would instead accumulate in the standard PC. With cross-content comparisons, the CI size becomes more uniform across different levels of quality.

All the results indicate that the scaling of the pairwise comparison data yields JOD scores that are highly correlated to MOS values acquired in the DSIS experiment. The introduction of cross-content pairs make JOD more uniform, and reduce the confidence intervals. Similar results have been showcased for other datasets, such as TID2013, in the paper [142].

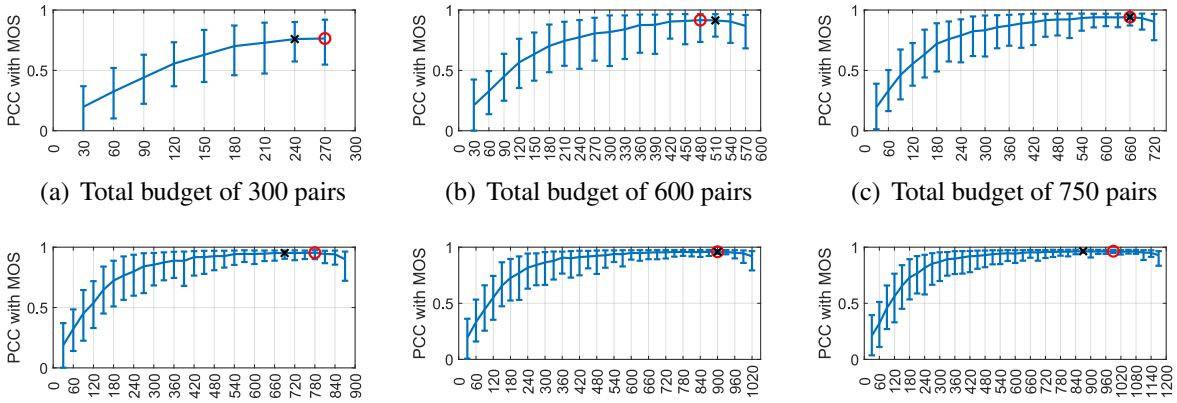
Table 2.4.1: Average confidence intervals of the videos with different bitrates (BR1 is the highest) for the considered experiments. The last column is the ratio of the CI of the combined PC data with additional cross-content pairs (CI_{CC}, CI of JOD_{CC}) to the CI of the combined PC data with additional same-content pairs (CI_{SC}, CI of JOD_{SC}). CI of standard PC experiment ($CI_{Standard}$, CI of $JOD_{Standard}$) are also reported for completeness.

Contents		$CI_{Standard}$	CI_{SC}	CI_{CC}	$Ratio_{CC/SC}$
Balloon	BR1	1.23	1.23	1.53	1.25
	BR2	2.21	1.68	1.86	1.11
	BR3	3.03	2.84	2.48	0.87
	BR4	3.93	3.36	2.56	0.76
Hurdles	BR1	1.45	1.50	1.12	0.75
	BR2	2.31	1.90	1.55	0.82
	BR3	3.12	2.36	2.46	1.04
	BR4	3.43	2.96	2.62	0.89
Starting	BR1	3.52	3.50	1.29	0.37
	BR2	4.45	4.06	1.47	0.36
	BR3	5.61	4.76	1.97	0.41
	BR4	6.04	5.11	2.29	0.45
Market	BR1	2.12	2.35	0.85	0.36
	BR2	3.05	2.80	1.63	0.58
	BR3	4.32	3.18	2.57	0.81
	BR4	4.73	3.28	2.94	0.90
Bistro	BR1	1.60	1.70	1.25	0.73
	BR2	2.12	1.91	1.46	0.76
	BR3	2.92	2.49	2.00	0.81
	BR4	3.34	2.91	2.26	0.78

What is the optimal fraction of cross-content comparisons?

In order to interpret the role of cross-content pairs in scaling, we can view pairwise comparison experiments as a graph, in which conditions represent nodes and comparisons edges. To scale the quality scores for such a graph in a consistent manner all conditions must be connected, i.e., there should be no disconnected components in the graph of comparisons. However, when each content is assessed individually, this forms a set of disconnected graphs, each with its own relative quality scale. We could potentially anchor each content by assuming that reference image for each content has a fixed quality score, for example, 0. However, then conditions far away in quality from the reference accumulate large measurement error. Thus, connecting these disconnected parts through cross-content comparisons is an essential step for unifying quality scale.

In this respect, when building this graph, i.e., deciding which conditions need to be compared, an interesting question is: *given a fixed budget of comparisons, how to allocate it among same-content and cross-content comparisons?* To answer this question, in [211] we have run sam-



(a) Total budget of 300 pairs (b) Total budget of 600 pairs (c) Total budget of 750 pairs

(d) Total budget of 900 pairs (e) Total budget of 1050 pairs (f) Total budget of 1200 pairs

Figure 2.4.4: The fixed-budget simulation results for HDRVDB. The x-axis indicates the number of same-content pairs for a total fixed-budget of N pairs, as N is indicated in the subcaptions. The whiskers indicate confidence interval for each case. The red circle indicates the case with the maximum correlation with MOS, and the black cross indicates the case with the minimum variation of PCC values.

pling simulations on two datasets. Specifically, we conduct a simulation in which we randomly sample a subset of same-content and cross-content pairs from the whole set of measured pairs (i.e., from the real experiment data), and we use this sub-sampled data to perform psychometric scaling. The sampling and simulation processes are described in [211].

We report here the results for the HDRVDB dataset in Figure 2.4.4, which shows the linear correlation coefficient between JOD's and MOS's as a function of the number of same-content comparisons. Notice that here the sum of same-content and cross-content pairs is constant. We also report the confidence intervals of PCC, estimated through bootstrapping. We observe that, in general, same-content pairs are more important than cross-content pairs, and we need a minimum number of same-content pairs before starting to add cross-content pairs. However, the best correlation with MOS is obtained when a small fraction of cross-content pairs, around 20% for this dataset, is considered. Cross-content pairs also help in stabilizing the results, minimizing the width of PCC confidence intervals.

Notice that these results are obtained by random subsampling of pairs, and thus all our conclusions must be interpreted in terms of the expected PCC with the (approximate) ground-truth quality scores. In other words, we do not take into account the impact of pair selection. The pair selection can have a great effect on the PWC scaling, and it can be done through, e.g., active sampling [203, 99].

2.4.4 Combining rating and pairwise comparisons

When results of both ranking and rating experiments are available for the same set of contents, it may be desirable to use all information when constructing the quality scale. In the following we propose a simple way of combining both types of measurements.

Based on the results of Section 2.4.3, we assume a linear relationship between random variables ω_i representing quality scores obtained from a pairwise comparison experiment (Eq. 2.4.2), and the random variables obtained from a rating experiment π_i:

$$\omega_i = a \cdot \pi_i + b. \tag{2.4.7}$$

We further assume that the standard deviation of the observer model may differ between both experimental protocols: people can confuse two conditions more often in one protocol than the other. Given that, the relationship is expanded to:

$$N(q_i, \sigma) = a \cdot N(m_{ik}, c \cdot \sigma) + b = N(a \cdot m_{ik} + b, a \cdot c \cdot \sigma), \tag{2.4.8}$$

where m_{ik} is the collected opinion score for the condition i and observer k. q_i is the latent quality score, which we want to recover. a, b and c are the unknown parameters that control the relationship between the rating and pairwise comparison data. Our goal is to find the values of the latent variables given the observed opinion scores m_{ik} and pairwise comparisons c_{ij}.

Since opinion scores are generally continuous, we express the probability of observing m_{ik} using the density function of the Normal distribution:

$$f(m_{ik}|q_i, a, b, c) = \frac{1}{\sqrt{2\pi a^2 c^2 \sigma^2}} e^{-\frac{((a \cdot m_{ik} + b) - q_i)^2}{2a^2 c^2 \sigma^2}}. \tag{2.4.9}$$

Assuming independence between observers, the likelihood of observing the whole set of opinion scores $\mathbf{M}$ is:

$$P(\mathbf{M}|\mathbf{q}, \sigma, a, b, c) = \prod_{i=1}^{N} \prod_{k=1}^{J} f(m_{ik}|q_i, \sigma, a, b, c). \tag{2.4.10}$$

Similarly, the likelihood of observing pairwise comparisons $P(\mathbf{C}|\mathbf{q}, \sigma)$ is given in Eq. (2.4.6). One advantage of this probabilistic formulation is that missing data, for example when observers rate only a portion of all conditions, can be simply omitted from the above product.

To recover latent quality scores $\mathbf{q}$ from both measurements, we use the maximum likelihood estimator with the posterior probability:

$$\arg\max_{\mathbf{q}, a, b, c} P(\mathbf{q}, a, b, c|\mathbf{C}, \mathbf{M}, \sigma), \tag{2.4.11}$$

where $P(\mathbf{q}, a, b, c|\mathbf{C}, \mathbf{M}, \sigma) \propto P(\mathbf{C}|\mathbf{q}, \sigma) \cdot P(\mathbf{M}|\mathbf{q}, \sigma, a, b, c) \cdot P(\mathbf{q})$ and $P(\mathbf{q})$ is a prior included to enforce convexity, e.g., uniform or Gaussian distribution.

Likelihood functions are scale-invariant, i.e. $P(\mathbf{M}|q, \sigma) = P(\mathbf{M}|tq, t\sigma)$ for a constant $t \neq 0$. Thus, without loss of generality, we can fix σ to an arbitrary value. As before, since scales are relative, we need to set an anchor, e.g. $q_1 = 0$.

Note that if we wish to mix different datasets, e.g. several datasets for which rating measurements have been collected, we can do so by collecting pairwise comparisons that link the data and running the optimization procedure previously presented. In this case, different standard deviation of the observer model and scaling parameters (a, b and c) should be assumed for different datasets. In particular, the value of c in (2.4.8) indicates the different degree of incertitude

in performing a rating assessment vs. a PWC on a given dataset. In [142] we performed mixed scaling and estimated the value of the parameter c for the TID2013 dataset, which we found to be 1.24. This suggest that in a typical image quality assessment experiment, the pairwise comparison protocol results in less confusion between observers.

Chapter 3

Image and video analysis

In many applications, images and video are not directly consumed by human observers, but are rather used as input to further processing aimed at extracting some *information* from them. This is the typical example of computer vision applications, aiming, e.g., at classifying objects or scenes; but also higher level tasks such as image understanding, or security applications where the goal is to assess the authenticity of a content. In this chapter, I review some of my past research activities on image and video analysis, which span several domains, from computer vision to forensics.

In Section 3.1, I consider the problem of image matching in presence of significant viewpoint changes. These represent one of the most challenging classes of geometric transformations for the basic block of image matching, which is local feature extraction. In our work, we show that having geometric information about the scene, e.g., in form of depth maps which are nowadays relatively easy to obtain, enables to design local features which are in principle invariant to changes of viewpoint of the camera. This is particularly interesting in scenarios such as for self-driving cars. In Section 3.2, I consider another class of transformations that affect matching performance: nonlinear illumination changes. In this case, we argue that high dynamic range imaging can be successfully used for countering even dramatic changes of illumination in a scene, without the need for designing new feature extraction pipelines. Specifically, we propose a learning-based framework to optimize a tone mapping operator for feature extraction, rather than for the traditional task of content display.

Section 3.3 moves from computer vision to the field of forensics and security, where the goal is to establish the authenticity of a content in a passive and blind manner. An essential problem in forensics is reconstructing the history of a content, e.g., in order to find its source. Here, we consider for the first time a forensic problem in the HDR domain: determining whether an HDR image has been obtained by multiple exposures or by single LDR image through inverse tone mapping.

Finally, I conclude the chapter with ongoing work on predicting the aesthetic quality of a picture in Section 3.4.

3.1 Local features for RGBD image matching

Local image features represent a key tool in a number of practical scenarios and applications in multimedia, including visual search, classification, indexing, image analysis, etc. The industrial demand for robust, distinctive and compact visual features has stimulated MPEG standardization activities for Compact Descriptors for Visual Search (CDVS) [54] and Compact Descriptors for Visual Analysis (CDVA) [55].

While 2D visual features have nowadays achieved a substantial level of maturity in terms of robustness, compactness, and efficiency, the emergence of richer image and video formats, such as texture+depth (RGBD), multiview or plenoptic images, have recently attracted attention towards the definition of features able to capture and leverage the *geometric* information of a scene [79, 70, 186]. Indeed, acquiring scene geometry is nowadays feasible with low-cost devices, such as Microsoft Kinect, which are capable of acquiring depth together with conventional color images.

The availability of geometrical information provided by depth could help to improve the performance of current image matching techniques in the presence of large variations of the camera viewpoint and *out-of-plane* rotations, where conventional feature schemes fail to detect and match repeatable keypoints.

In this section I describe several contributions aimed at employing the depth information in order to extract repeatable keypoints and describe them in a way that is as invariant as possible to the camera position. I start by introducing the basic concepts of image matching and invariance in Section 3.1.1. In Section 3.1.2 I describe how to extract viewpoint and scale-invariant keypoints by using a smoothing operator that engenders a scale space in the RGBD domain [72, 74]. Finally, in Section 3.1.3, I will describe a complete local feature extraction pipeline for RGBD content matching.

3.1.1 Background concepts on image matching

The problem of finding local correspondences between images is a fundamental task in vision. The common framework to solve this problem is referred to as *image matching*. It consists of three main stages.

1. **Detection**: each input image is processed independently to find repeatable salient visual points.
2. **Description**: a compact signature representing a neighborhood of each detected point is computed.
3. **Matching**: signature sets from different images are compared, producing a set of correspondences.

Salient visual points are also called *keypoints*. Typical keypoints could be corners or blobs (regions with similar characteristics) in the image. In order to design a local feature extraction pipeline, we thus need to design a *detector* and a *descriptor*.

The common purpose of image matching is to recognize (semantically) the same content in different acquisition conditions. In other words, the two images being matched typically represent the same or similar content, and are related by a ***visual deformation***. From this perspective, two key concepts may be defined: *covariance* and *invariance*. In order to be able to match the same content within the two input images, we expect to extract the same or very similar descriptors. Therefore, when the observed content undergoes a deformation, the descriptors are expected to remain the same or ***invariant***, so that they provide a representation of the content and not conditions of its acquisition. Contrary to the descriptors, the keypoints are expected to be ***covariant***, i.e., when a deformation occurs, they are expected to follow it (change accordingly). The keypoints thus depend on the deformation and represent the conditions of acquisition rather than the content itself. Together these two concepts are generalized to ***feature stability***: a *stable feature* is such a feature that allows to match two images related by a corresponding deformation, i.e., it remains detectable when the content undergoes this deformation. To be stable, a feature needs a covariant keypoint and an invariant descriptor. By convention, we sometimes say that a feature is *invariant* when it is stable to a given visual deformation. Another commonly accepted term referring this quality is *feature repeatability*.

The degree of feature stability may be qualitatively measured by the nature of visual deformations the given feature is robust to. A simple classification of the deformations affecting feature stability is given in Table 3.1.1 [78]. To some of the listed deformation classes, e.g., image noise (**P-III**), no feature could be perfectly stable neither totally unstable: one can measure the stability quantitatively, for example, adding a progressively increasing noise to the image and trying to match it against its noiseless original. However, to the most part of other deformations, notably geometric ones, a given feature may be *invariant by design*. For example, many existing local features in traditional imaging are invariant by design to the first three classes describing orthogonal transformations in camera plane. Also, some simple illumination changes (**P-I**) (such as affine $I \rightarrow \alpha I + \beta$ for α and β constant all over the image I) are typically covered too. A classic example of translation, rotation, scale and (partially) illumination invariant feature is SIFT [104].

To be invariant by design to a specific deformation class, a feature extraction process must involve processing techniques that are themselves covariant and invariant to that class. For example, in-plane scale changes are handled by involving a multiscale representation on the detection stage that allows to discover scale-covariant keypoints, and the descriptor patch is then scaled accordingly to the detected *characteristic scale*.

In many application scenarios exploiting image matching as a basic task, the observer and/or the objects can move arbitrarily not only in the camera plane, but in all the three dimensions. This causes *perspective distortions*. In the context of local features, they are often seen as an effect of ***out-of-plane rotations*** (**G-IV**). Due to the locality of the features, these deformations become equivalent to unconstrained *local tridimensional rigid deformations* of the observed content, which is arguably the most common kind of visual distortions in practice. Invariance *by design* to out-of-plane rotations is unlikely to be achieved using only photometric information. This reveals a weak point of the paradigm of photometric local features. For example, SIFT may demonstrate limited performance when the content undergoes out-of-plane rotations more than 40° [104, 126]. This problem may be addressed when the image is complemented by a geometry description. This is the focus of the rest of this section.

Table 3.1.1: A classification of the most common visual deformation classes in the context of feature matching robustness. The degree of stability of a given feature extraction approach can be assessed by its capacity to perform well when a deformation of the corresponding class is present between the two matched images. We denote different classes with "G" for geometric deformations and "P" for photometric ones, ordering them in each group by arguably increasing complexity from the feature matching points of view. Courtesy from [78].

G-I	In-plane translations	Geometric	Rigid
G-II	In-plane rotations		
G-III	Scale changes		
G-IV	Out-of-plane rotations		
G-V	Affine deformations		Non-rigid
G-VI	Isometric deformations		
G-VII	Non-isometric deformations		
P-I	Affine illumination changes	Photometric	
P-II	Non-linear illumination changes		
P-III	Image noise		

3.1.2 Keypoint extraction based on a RGBD scale space

In this section we focus on the first step of the feature extraction pipeline, i.e., *keypoint detection*. A texture+depth (RGBD) image could be considered as a mesh with an associated texture. Thus, RGBD matching could be cast as a problem of mesh matching, where several techniques have been proposed in the literature [204, 200, 87]. However, these techniques are not apt to deal with occlusions, that are commonly present in images, and the repeatability of detected keypoints is intrinsically limited by resampling when the camera moves. Therefore, image-level techniques for feature detection on RGBD content are of interest. The proposed approach consists in exploiting the depth map in order to define a non-uniform scale space for the texture image. The process we define aims at exploiting the surface properties that do not depend on the observer position in order to render a viewpoint-covariant multiscale representation that is able to reveal robust keypoints. The construction of the scale space is summarized below, and described with more mathematical details in [74].

Definition of RGBD scale space

In order to construct the proposed RGBD scale space, we first need to define a Laplacian operator, such that it enables to establish a diffusion process. The first step is to define a parametrization of the image surface in local camera coordinates as illustrated in Fig. 3.1.1:

$$\vec{r}(u,v) = \begin{pmatrix} 2u\tan\frac{\omega}{2} \\ 2v\frac{H}{W}\tan\frac{\omega}{2} \\ 1 \end{pmatrix} D(u,v). \tag{3.1.1}$$

Based on this reparametrization, we can define first-order differential quantities, which are similar to directional derivatives, e.g.:

$$\partial_u f = \frac{f(u+h,v) - f(u-h,v)}{\|\vec{r}(u+h,v) - \vec{r}(u-h,v)\|}, \tag{3.1.2}$$

and similarly for $\partial_v f$. Second-order differential quantities are obtained by applying twice these operators, e.g., $\partial_{uu} f = \partial_u(\partial_u f)$. A more precise formulation of these quantities is available in [74]. Finally, we define a Laplacian-like second order differential operator summing up the second-order differential quantities defined above:

$$L \equiv \partial_{uu} + \partial_{vv}. \tag{3.1.3}$$

Next, we set up a partial differential equation problem that describes the diffusion process with the proposed Laplacian operator (3.1.3):

$$\begin{cases} \dfrac{\partial f}{\partial t} = Lf \\ f|_{t=0} = f_0. \end{cases} \tag{3.1.4}$$

This problem is very similar to the classic diffusion problem. To study this similarity we consider the continuous case of this problem. We obtain a continuous generalization of the differential quantities defined above by letting h tend towards zero, that is:

$$\begin{aligned} \mathcal{D}_u f &= f_u \|\vec{r}_u\|^{-1} \\ \mathcal{D}_{uu} f &= f_{uu}\|\vec{r}_u\|^{-2} - f_u\|\vec{r}_u\|^{-4}(\vec{r}_u, \vec{r}_{uu}). \end{aligned} \tag{3.1.5}$$

Thus, we get the continuous version of problem (3.1.4):

$$\begin{cases} \dfrac{\partial f}{\partial t} = \mathcal{D}_{uu} f + \mathcal{D}_{vv} f \\ f|_{t=0} = f_0. \end{cases} \tag{3.1.6}$$

It is worth noticing that if the depth D is constant (i.e., we have a non-informative depth map), this PDE problem becomes equivalent to the classic linear diffusion filtering, as the differential operator on the right side of the equation turns into the classic Laplacian up to a constant multiplier due to $\vec{r}_u = \vec{r}_v \equiv const$ and $\vec{r}_{uu} = \vec{r}_{vv} \equiv 0$. This allows for a "backward compatibility" of the

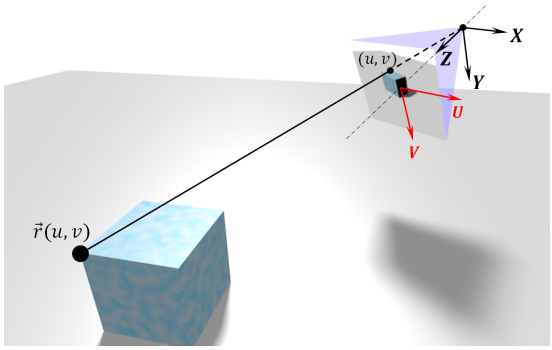


Figure 3.1.1: Scene surface parametrization in local camera coordinates.

proposed scale space to the classic Gaussian scale space in the case when the depth map is not provided. Moreover, this property is satisfied locally, i.e., at points where D is continuous and the surface normal is parallel to the camera optical axis. It can be shown that Problem (3.1.5) is well posed, causal (no spurious features will appear during the smoothing process), and is numerically stable, and can be efficiently solved on GPU's [74].

The designed filter simulates a uniform smoothing along the scene surface through a non-uniform diffusion in the image plane. Since smoothing along surfaces is, in principle, independent on the observer position, the proposed scale space can provide keypoints that are invariant to viewpoint position changes. This behavior is referred to as *viewpoint covariance*, as discussed in Section 3.1.1. It mainly comes from the definition of the first order differential operators (3.1.2), where we weight the derivative computed on two neighboring samples by the real distance between the corresponding sample points on the scene surfaces, inferred from the depth map. In practice, this diffusion process only approximates a diffusion process on the manifold defined by the depth map, due to depth errors and texture sampling precision. Therefore, the resulting scale space behavior will be approximately viewpoint covariant.

Some examples of images obtained with the proposed smoothing operator compared to the Gaussian smoothing are presented in Fig. 3.1.2. The input image is taken from the LIVE dataset [179, 178], which provides depth maps captured through a laser scanner. The viewpoint-covariant behavior could be observed on large scales (images (b), (c), (e), (f)): as the smoothing is propagating along the surface, and not uniformly in the image plane (as in case of the Gaussian scale space), the image becomes less smoothed when the distance increases.

Proposed detector

A keypoint detector mainly consists of three parts: (i) initial keypoint candidates selection criteria selecting a set of locations with corresponding scales in the input image, (ii) a candidate filtering, aimed at rejecting candidates that are likely less repeatable, and (iii) an accurate localization procedure of remaining keypoints.

Similarly to the popular SIFT detector [104], the initial keypoint candidates in our proposed detector are selected as local extrema of the Laplacian operator (3.1.3). the main difference with respect to the SIFT detection criterion is that *we look only for spatial local extrema at each scale*,

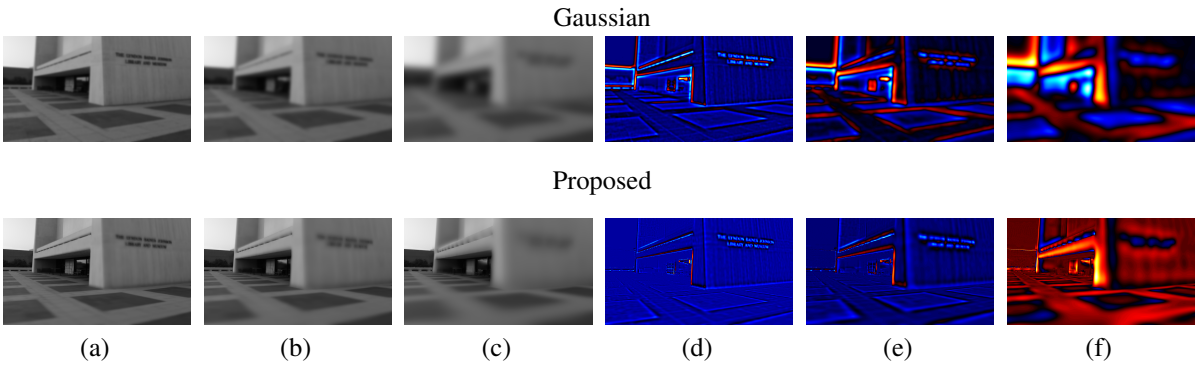


Figure 3.1.2: An example of the proposed scale space on a real RGBD image. Top row: standard Gaussian scale space (no depth map used), second row: the proposed scale space. Images (a), (b) and (c) in each row present different levels of smoothing: $\sigma = 5$, 10 and 25 for the Gaussian scale space and $\sigma = 0.1$, 0.2 and 0.5 for the proposed one. Images (d), (e) and (f) represent corresponding Laplacian operator outputs.

i.e., over variables u, v, and not for the local extrema along both spatial and scale coordinates, i.e., over u, v and σ. Indeed, in our experiments we found that keypoint candidates issued from extrema along the σ axis are generally unstable. A possible reason for that is related to the intrinsic nature of our proposed scale space: the smoothing injected into the image is spatially varying, so that σ represents a scale with respect to the scene geometry, and not the scale in the image plane. On the other hand, local minima and maxima of our Laplacian (3.1.3) with respect only to spatial image variables u, v turn out to be very repeatable, and reveal distinctive blob-like structures on the scene surface.

Specifically, we construct a set of smoothed images of levels $\sigma_0, 2\sigma_0, 4\sigma_0, ..., 2^{M-1}\sigma_0$ ($M = 5$ in our experiments). Here σ_0 is a constant, i.e., its value is set manually according to the depth measurement unit used in the depth map. Each subsequent image is subsampled by two in each dimension with respect to the previous one. After selecting local extrema, we filter out the keypoints localized on the edges that are likely to be unstable, as they can move along the edge when the camera position changes. In order to localize keypoints with subsample precision, we apply the accurate localization procedure presented in [17], reducing it from three dimensions (u, v, σ) to two (see [74] for details).

After the keypoints are detected, in order to be able to use standard descriptors, we derive their on-screen scale. We consider keypoint k as a sphere of radius σ_k, situated on the scene surface. σ_k is simply equal to the scale level where the keypoint is detected. Assuming that its center is projected on the screen at point (u_k, v_k), obtained from the accurate localization procedure, we apply the pinhole camera model to get the output (on-screen) keypoint scale:

$$s_k = \frac{\sigma_k W}{2D(u_k, v_k) \tan \frac{\omega}{2}}. \tag{3.1.7}$$

The set of triples $\{(u_k, v_k, s_k)\}_k$ constitutes the detector output and is sent to the descriptor extraction stage. An example of detected keypoints in an image from *Bricks* sequence is given in Fig. 3.1.3.

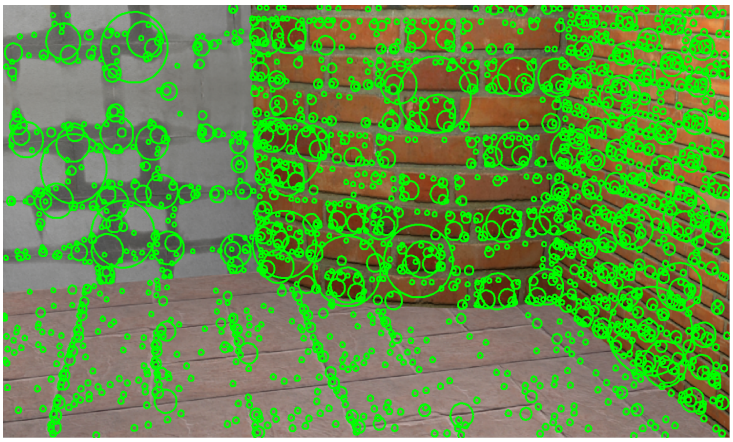


Figure 3.1.3: Keypoints detected using the proposed method in an image of *Bricks* sequence.

Repeatability evaluation

Repeatability [118, 119] is a commonly used measure to evaluate a keypoint detector. The evaluation consists in extracting keypoints from several images (views) of a given scene, and then counting the portion of repeated keypoints between a reference view and each remaining view. The keypoint A coming from the reference view is considered as repeated if there is a keypoint B in the test view that covers (approximatively) the same area of the scene. An *overlap error threshold* $\eta \in (0,1)$ is employed to decide how strict the correspondence between keypoints should be: the smaller η is, the more precisely the keypoints should be repeated. The *repeatability score* is the number of repeated keypoints divided by the maximum possible number of repetitions. For the latter we take the maximum number of keypoints detected in one of the two views, excluding those keypoints that fall out of the field of view of any of the two cameras, so that only the surface area present in both views is considered.

We compare the proposed detector to the standard SIFT detector and to Viewpoint Invariant Patches [201], which incorporates a keypoint detector that uses the depth map. Three RGBD test sequences are used [71], representing different content, containing significant viewpoint position changes: *Bricks* (20 images), *Graffiti* (25 images, re-synthesized from the original *Graffiti* sequence from [118]) and *House* (25 images). The repeatability score of each detector is computed for two values of the overlap error threshold $\eta = 0.5$ and $\eta = 0.25$. The results of this experiments are shown in Fig. 3.1.4. It can be observed that, for both values of the overlap η, the proposed detector clearly outperforms the two other approaches. Moreover, even in the tighter condition $\eta = 0.25$ our proposed detector demonstrates a comparable or better repeatability to the two other detectors, even when those are matched using the more tolerant value $\eta = 0.5$.

Scene recognition using Kinect images

The proposed RGBD detector can be employed in a simple scene recognition application which requires repeatable local features. The application scenario is, e.g., for a mobile robot or a drone, to recognize the location (room) where it is situated, solely using visual sensors data and prior

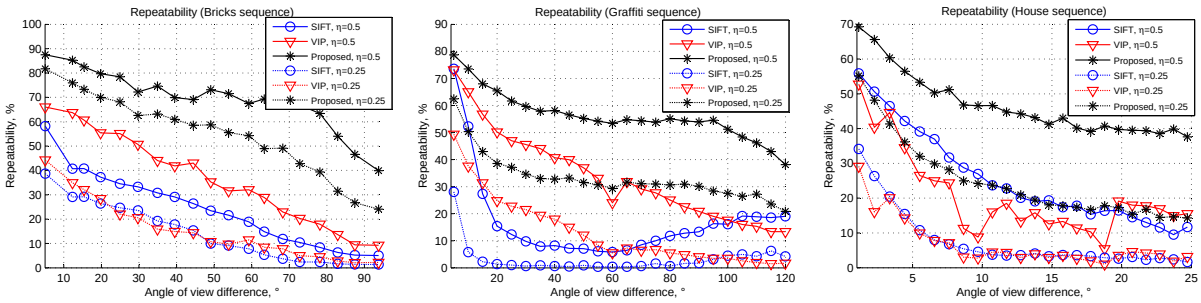


Figure 3.1.4: Repeatability score on synthetic RGBD sequences in function of angle of view difference between reference and test images.

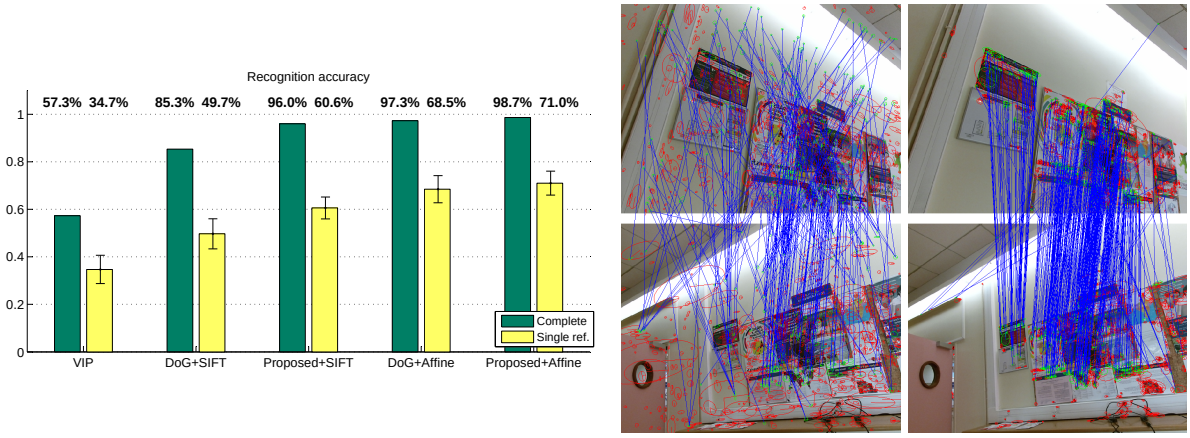
knowledge, i.e., a database of local features representing different locations. Using Microsoft Kinect sensor, we captured 75 RGBD images in 15 different indoor location (5 images per location taken from different positions, but in such a way that the same objects are visible in all the 5 images).

To this end, we detect keypoints in a query image and in each image of the dataset, and match their corresponding descriptors. We use state-of-the-art descriptors jointly with the newly proposed detector, in particular:

- original VIP features [201],
- standard SIFT features (*VLFeat* [193] implementation, referred to as DoG+SIFT),
- SIFT descriptors undergoing affine normalization [117], bootstrapped with SIFT keypoints (*VLFeat* implementation, referred to as DoG+Affine),
- our proposed detector with standard SIFT descriptors (referred to as Proposed+SIFT),
- SIFT descriptors undergoing affine normalization [117], bootstrapped with our proposed detector (referred to as Proposed+Affine).

If the number of closely matching descriptors is large enough, then the two images are assumed visually similar, and the location of the image is assigned consequently. Details on the experimental protocol are available in [74]. The portion of correctly classified images per method in this setting is reported in Fig. 3.5(a), for two classificatio scenarios (see the description in the figure). Our proposed detector achieves a higher recognition accuracy in both the experiments. Affine normalization compensates the perspective distortions on the descriptor computation stage, yielding improved performance compared to the unnormalized SIFT descriptors.

For qualitative comparison, an additional illustration of matching using these descriptors is given in Fig. 3.5(b): keypoints detected with the proposed detector generally provide more consistent and regular correspondences. Moreover, in spite of the noise present in depth maps and their incompleteness (some areas have undefined depth, which is a common problem of infrared depth sensors), our proposed approach is able to detect repeatable keypoints.



(a) (b)

Figure 3.1.5: (a) Accuracy of scene recognition. The left bars (*complete*) are computed by matching a query image to all the remaining 74 images in the dataset. In the *single-reference* classification, instead, each image is classified using a set of 15 randomly selected reference images (one per class). In this case the reported results are the average over 100 repetitions, and the error bar is the corresponding standard deviation. (b) Raw (putative) feature matches between two RGBD images from *Board* scene obtained with affine-covariant descriptors on top 1000 keypoints in each image. Left: SIFT detector (243 matches), right: the proposed detector (419 matches).

3.1.3 TRISK: local features extraction for RGBD content matching

In this section we consider a different kind of keypoints commonly used in computer vision: *corner points*. Differently from Section 3.1.2, here we design a *complete* feature extraction pipeline consisting of a detector and a descriptor [77]. The key idea consists in involving the surface metric, derived from the depth map, into the texture map processing, in the form of *adaptive local axes* replacing the regular image coordinates. This allows to apply Accelerated Segment Test (AST) in an intrinsic way to the scene surface and render keypoints more stable. This test lies at the basis of the proposed detector. The proposed descriptor is based on an efficient local planar normalization resulting from the local axes.

We build the proposed scheme following the *binary local features* paradigm. One of the first proposed binary features, BRIEF (Binary Robust Independent Elementary Feature) [19] extends the idea of local binary patterns [136], originally designed for texture analysis tasks, to describe interesting points. Since the extracted feature is a string of bits, the matching is done using Hamming distance, which is more efficient to compute than the Euclidean one. This idea is further elaborated in numerous works [165, 97, 3]. Notably, ORB (Oriented FAST and Rotated BRIEF) [165] and BRISK (Binary Robust Invariant Scalable Keypoints) [97] present complete extractors of scale and rotation invariant binary features. They apply *FAST* [164] and *AGAST* [109] corner detectors to scale space-like image pyramids to find the keypoints, estimate dominant keypoint orientations, and then invoke the same principle of binary description. The feature proposed in this work employs a similar binary pattern, but we sample it in the scene surface rather than in the camera plane. To underline the continuity with the visual feature liter-

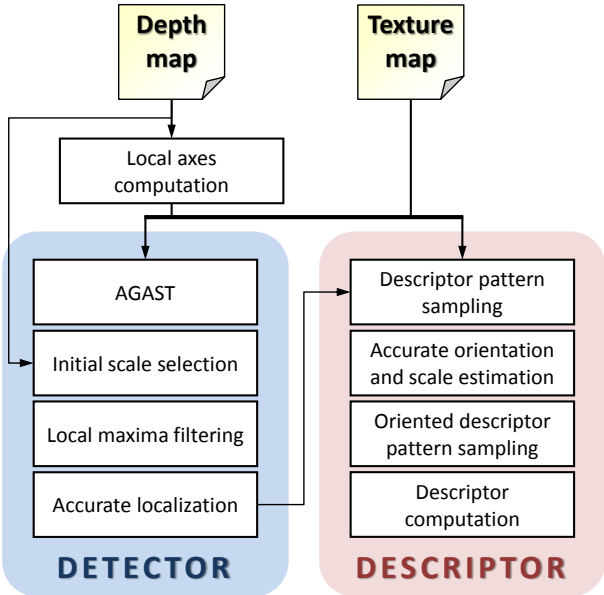


Figure 3.1.6: The architecture of the proposed TRISK pipeline. TRISK is a complete feature extraction framework for RGBD content, composed by a keypoint detector and a descriptor. Both leverage the geometric information provided by the depth map in order to sample the texture considering a different local coordinate system for each point of an object surface. The *detector* is based on the Adaptive Generic Accelerated Segment Test (AGAST) response, computed in local coordinates. Depth is also used to find the approximate *geometric* scale of a keypoint, which is further refined at the description stage together with orientation normalization. The local maxima filtering and accurate localization stages enable to select the most repeatable keypoints. In order to compute the *descriptor*, the texture is sampled again in local coordinates. A multi-pass procedure is employed to accurately estimate the orientation and scale of the sampling pattern. Finally, similarly to the BRISK descriptor, pairwise comparison tests across the texture samples are carried out to produce a binary descriptor string.

ature and specifically BRISK, we then call the proposed features **TRISK**, for "Tridimensional Rotational Invariant Surface Keypoints". The overall scheme of TRISK is shown in Fig. 3.1.6.

Local axes computation

In order to render the feature extraction process as independent as possible of the camera position, we adapt all the local processing to the surface geometry, considering the observed image as a textured manifold. In TRISK, we follow this way by selecting a proper basis at each image point, which we further refer to as *adaptive local axes*. They are used to transfer the detection and the description from the camera plane onto the scene surface, basing them on the surface metric, which is intrinsically independent of the reciprocal camera-to-object position and orientation.

Assuming that keypoint detection and description are rotationally invariant, the local axes are given by any orthonormal basis of the tangent plane, projected on the camera plane and normalized so that its largest vector has unit norm in pixels. An intuitive geometrical explanation

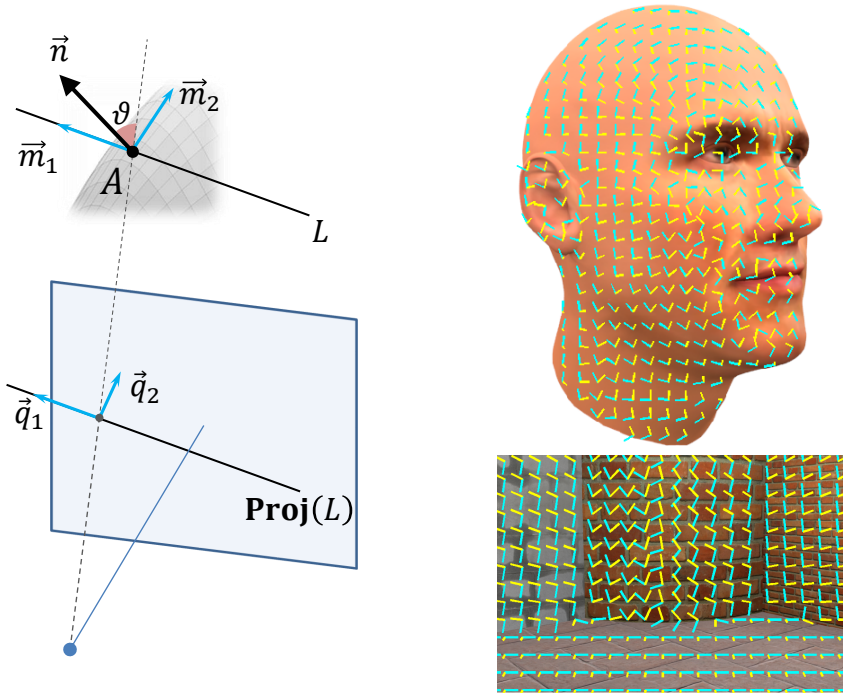


Figure 3.1.7: Computation of local axes $\vec{q}_1$ and $\vec{q}_2$. On the left: $\vec{q}_1$ and $\vec{q}_2$ are obtained by projecting $\vec{m}_1$ and $\vec{m}_2$ in the 3D space onto the camera plane. $\vec{m}_1$ is chosen to be always parallel to the camera plane, and its projected local axis is normalized to unit length. The projection of $\vec{m}_2$, i.e., $\vec{q}_2$ has a length reflecting the perspective distortion at A, which depends on the angle ϑ between the viewpoint $\vec{A}$ and the normal at A. On the right: examples of local axes fields computed on images from *Arnold* and *Bricks* sequences, with $\vec{q}_1$ shown in cyan and $\vec{q}_2$ in yellow.

and examples showing local axes $(\vec{q}_1, \vec{q}_2)$ for two RGBD scenes are shown in Fig. 3.1.7. The local axis field can be computed based on the depth information, based on the projection model we introduced in Figure 3.1.1. In particular, the local axes $\vec{q}_1$ and $\vec{q}_2$ depend only on the surface normal and the point position on the camera plane (u, v), but not on the depth map values directly. To estimate the normal vector we use PCA-based normal estimation [166]. Therefore, differently from the scale space introduced in Section 3.1.2, TRISK avoids explicit manipulations with differential characteristics of the depth map, which are prone to noise. In [77] we derive an analytic expression of the local axes field, allowing to compute them efficiently at each pixel location.

Detector

We employ the Adaptive Generic Accelerated Segment Test [109] to detect corners in the local adaptive axes, as illustrated in Figure 3.1.8. According to this test, a pixel is deemed to be a corner if it is darker or brighter than at least N connected points on a circle $\{(u_k, v_k)\}_{k=1}^{N}$ surrounding it. In TRISK, the texture map is interpolated using the local surface axes described above. Specifically, in order to transform the circle into its equivalent $\{(x_k, y_k)\}_{k=1}^{16}$ on the local

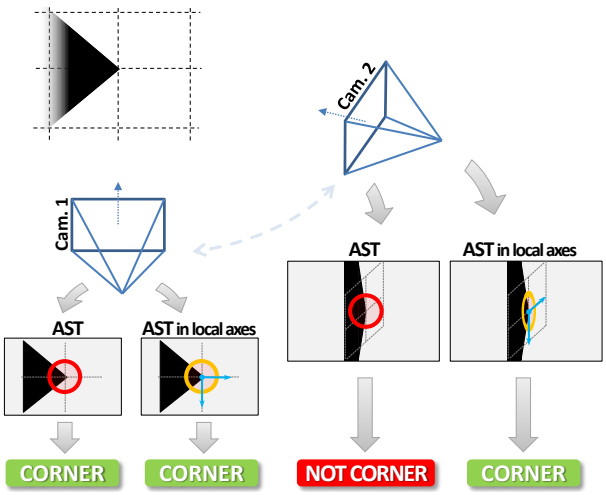


Figure 3.1.8: Illustration of application of Accelerated Segment Test (AST) in standard image axis versus local axes derived from the depth map. A corner viewed under a large angle projects itself at a nearly straight contour on the camera plane, so that the corner test in standard image axes fails causing a repeatability loss.

axes $(\vec{q}_1, \vec{q}_2)$, we need to perform a change of basis, i.e., we sample the texture map at locations

$$(x_k, y_k) = (u_k\xi_1 + v_k\xi_2, u_k\eta_1 + v_k\eta_2), \quad k = 1, \ldots, N. \tag{3.1.8}$$

The corner test is then performed on the obtained samples. The idea of performing AGAST in local axes is illustrated in Fig. 3.1.8. Non-local maxima suppression is then applied on the generated score map in order to select the keypoint candidates.

To derive the keypoint scale we exploit the depth map similarly to [28]. We use AGAST response only to derive the keypoint position but not its scale, since in case of RGBD images a better clue of scale is available in the depth map. To achieve scale invariance, we employ the *geometrical scale*. Namely, we get the keypoint scale from the depth map assuming that the underlying visual detail is of a fixed spatial size σ_0. Keypoints farther from the camera have smaller spatial support in pixel units, due to perspective distortion. σ_0 is the coefficient of this inverse proportionality relation, which defines a sort of "anchor" size to which objects (in spatial units measured in the camera plane) are scaled based on their depth. That is, $\sigma = \frac{\sigma_0}{z}$, where z is the average depth of the keypoint. Intuitively, σ_0 is related to the characteristic size of repeatable landmarks, which depends on the content and viewing conditions. The optimal value of σ_0 is found by grid search as explained in [77]. The keypoint area is finally described by an ellipse spanning the scaled local axes $\sigma\vec{q}_1$ and $\sigma\vec{q}_2$. Thus, TRISK keypoints are not circular as those of SIFT or BRISK, or those obtained by our RGBD scale space shown in Figure 3.1.3; they are rather elliptical, similarly to the keypoints produced by affine-covariant detectors [117].

Local maxima filtering to select the most stable keypoint candidates, and accurate localization of keypoint position are carried out similarly to conventional local features such as SIFT, and are detailed in [77].

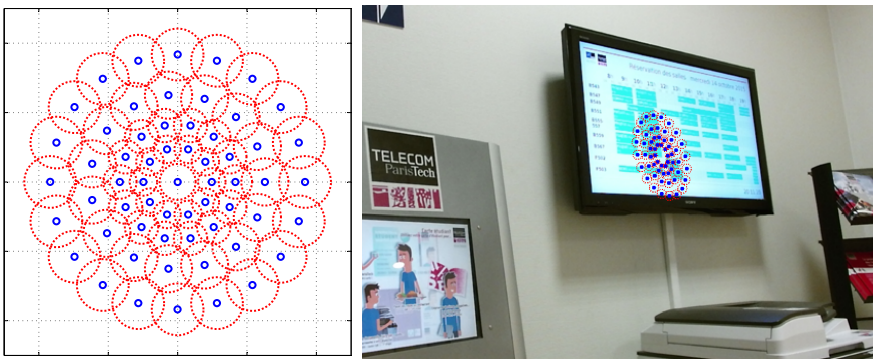


Figure 3.1.9: BRISK descriptor sampling pattern from the original implementation (left) and its mapping to the surface through local planar normalization (right).

Descriptor

Once the set of interesting point positions and scales is provided, a compact description is computed for each point. In [71], we studied how binary features may be used to extract a surface-intrinsic information from RGBD images in order to provide a description robust to rigid 3D deformations. A descriptor sampling pattern was projected on the scene surface, providing a depth-based descriptor normalization procedure aimed at producing invariant features. However, such a projection is (1) very sensitive to depth map noise and (2) requires a high computational effort. To be robust to the viewpoint position changes on the descriptor level, in TRISK we propose a simpler approach based on a similar concept: the descriptor normalization is performed according to the local tangent plane approximating the scene geometry nearby the keypoint, computed directly in the camera coordinates using the definition of local axes.

Specifically, we reuse the BRISK descriptor sampling pattern, which consists in computing the average image intensity at certain points and within a given range, as shown in Figure 3.1.9. We apply this pattern to the image in adaptive local axes computed at the keypoint, that immediately gives us the approximating local plane. An example of how the BRISK pattern is mapped onto the scene using local axes at a given corner point is shown in Fig. 3.1.9. We notice that our design is not restricted to the BRISK sampling pattern; another manually designed or appropriately learned pattern, e.g. [3] or [165], might be used with no additional cost. In [77], we propose a three-pass sampling scheme that estimates accurately both the dominant orientation and scale.

Experimental results

We compare the performance of TRISK with 5 state-of-the-art features, summarized in Table 3.1.2. We test these methods on the RGBD sequences used in Section 3.1.2 and on the *Freiburg* RGBD dataset [177].

As a first evaluation on the syntehtic RGBD sequences, we compute the matching score and the ROC curve for each feature, by matching the first image of the sequence to the remaining ones. In a nutshell, matching score allows to judge on the ability of the detector to produce repeatable

Table 3.1.2: Summary of compared methods.

Method	Keypoint type	Descriptor type and size	Depth map use
TRISK	Corner	Binary 512 bit	detector and descriptor
BRISK [97]	Corner	Binary 512 bit	no
STAR-BRAND [28]	Blob	Binary 512 bit	descriptor
VIP [201]	Blob	Numeric 128 dim.	preprocessing
AFFINE [117]	Blob	Numeric 128 dim.	no
SIFT [104]	Blob	Numeric 128 dim.	no

keypoints as well as on the matching capability of the entire pipeline, whereas ROC shows how the descriptors are discriminative, e.g., their ability of distinguishing salient visual information in presence of deformations. Put together, these characteristics trace the two main axes of the local visual features mid-level evaluation: *repeatability* and *distinctiveness*.

More specifically, to compute matching score we first match each descriptor in the reference image to the descriptors in the test images, by computing an inter-descriptor similarity measure (Hamming or Euclidean distance, depending on the descriptor type). Two descriptors are then said to match if their distance is below a given threshold (*putative matches*). The set of putative matches between the two given images is split into correct (*true positive*) and incorrect (*false positive*) matches using ground truth. Two keypoints coming from different images but occupying the same area of the scene are called *repeated keypoints* (see Section 3.1.2); they produce a correct match if the descriptors corresponding to these keypoints are matched. The ratio between the number of correct matches and the maximum possible number of matches is the *matching score* per image pair. By varying the threshold to decide putative matches, one can compute the true and false positive rates and trace the ROC curve.

The resulting matching score and ROC curves obtained on the synthetic RGBD test sequences are presented in Fig. 3.1.10 — similar results for the *Freiburg* dataset are available in [77] and are not reported here for the sake of space. It can be seen from the results that in all the test sequences TRISK demonstrates improved overall matching score. In some cases (*Graffiti*, *House*) TRISK also shows the slowest decay, which indicates improved feature stability under viewpoint position changes. The second best matching score on synthetic sequences (top row in Fig. 3.1.10) is arguably achieved by VIP. Based on a planar normalization technique, VIP performs well in case of simple geometry, i.e., when the scene surface is mostly planar or very smooth, otherwise it may even be unable to detect any features. TRISK also exploits the principle of planar normalization, but in a much more local way, which allows it to perform well in scenes with more complex geometry, such as *House*.

In terms of ROC curves, we observe that TRISK performs relatively well but it is sometimes outperformed by non-binary features like SIFT (which use a more precise numeric representation for descriptors, see Table 3.1.2), and VIP. For *House* the latter has apparently a higher discriminability, but its matching score drops dramatically after 10°of out-of-plane rotation. Descriptors can actually be fairly compared only for similar values of matching score.

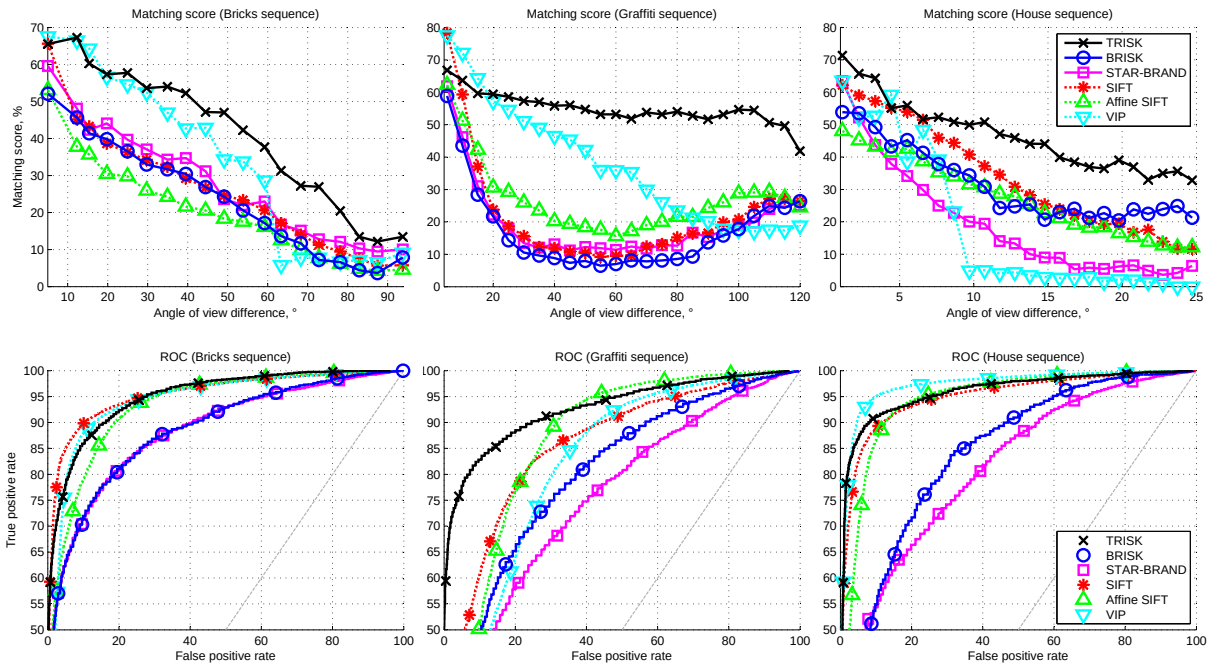


Figure 3.1.10: Matching score and receiver operating characteristics demonstrating repeatability and distinctiveness of the compared detectors and descriptors, mainly under out-of-plane rotations (*Bricks* and *Floor* sequences) and scale changes (*House* sequence). Computed on synthetic RGB data.

We also assess TRISK performance in a visual odometry scenario using image sequences from the *Freiburg* dataset [177]. The goal consists in retrieving camera pose evolution relatively to an initial pose using only the acquired images. The ground truth pose is recorded with a motion capture system and is provided within the dataset. We follow the setting of [28]: to compute the camera transformation (translation and rotation) between two frames, we match them, apply RANSAC to filter putative matches and, finally, run the Iterative Closest Point algorithm [9] retrieving the relative translation vector and rotation matrix. The resulting pose is recovered by cumulating deduced translations and rotations. To perform the evaluation, we subsample the sequence temporally in order to provide more challenging matching conditions, and we evaluate the translation and rotation error with respect to the ground-truth camera position and orientation.

The evolution of these errors on a sequence from *Freiburg* dataset is presented in Fig. 3.1.11 — other examples are reported in [77]. We observe that TRISK generally achieves smaller and bounded errors, in particular on the estimation of the rotation, where the approximate invariance to viewpoint changes makes the proposed feature more robust. It is worth noticing that on this sequence VIP is unable to provide enough matches for continuous trajectory estimation, and thus is not reported.

We conclude these results by observing that TRISK could be improved in its ability to deal with complex, highly detailed geometry, currently limited by the local planar approximation used to compute the descriptor. We have proposed a more complex way to render the descriptor stable and invariant to viewpoint position changes in [71]. However, this is more computationally

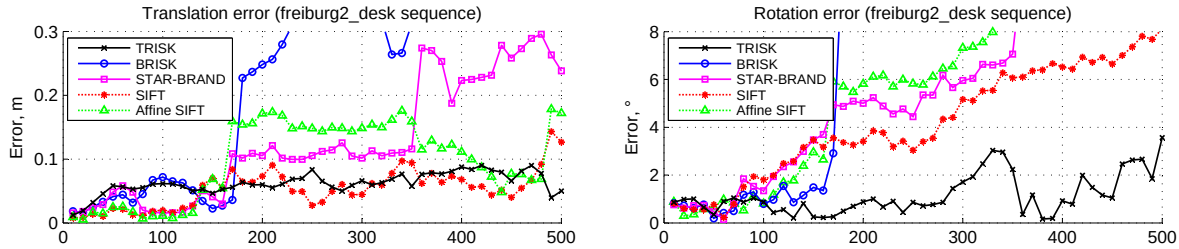


Figure 3.1.11: Visual odometry with 10 frames skipping on *freiburg2_desk* sequence (first 500 frames): translation (top) and rotation (bottom) errors. VIP fails on this sequence, thus it is not reported.

expensive and sensible to the depth map noise. Rendering the descriptor robust to geometrically complex scenes is still an open problem, although recently approaches based on convolutional neural networks have been showing interesting results [205].

3.2 Learning-based tone mapping for robust image matching under illumination changes

In contrast to the previous section focusing on robustness to out-of-plane rotations, in the following I will consider another class of challenging transformations for image matching: that of non-linear illumination changes (class **P-II** in Table 3.1.1). The performance of computer vision algorithms can substantially degrade with drastic lighting variations [214]. A possible solution to this problem consists in using a contrast-preserving acquisition technology such as high dynamic range imaging. HDR imaging could in principle produce features invariant to photometric changes, as it enables to draw on subtle, yet discriminating details present both in the extremely dark and bright areas of a scene, which would otherwise get lost.

However, even when an HDR acquisition of the scene is feasible, matching HDR images directly does not guarantee good matching results. We have shown in [151, 152] that using HDR linear (radiance) values significantly biases the localization of keypoints towards the extremely bright areas, leading to poor matching performance using conventional feature extraction pipelines. Differently from the previous section where we designed new detectors/descriptors for RGBD, in this work we aim at using existing, state-of-the-art local features. Therefore, we propose to modify the *input* to these feature extractors, i.e., the HDR images, by applying a proper tone mapping operator (TMO). Specifically, based on the observation that HDR images need to be used by a computer vision algorithm and not displayed to a human observer, we *design a TMO that is optimal for the specific task of feature extraction*, rather than for display purposes. In this respect, this work shares similarities with the TMO optimization for video compression described in Section 1.2, in that it is task specific rather than oriented to visual perception.

Optimizing a TMO considering keypoint detection and description concurrently is not trivial, as the corresponding design objectives are generally different. In this section, we address this problem and design an **optimal tone mapping operator (OpTMO)** to enhance the detection

and matching of features extracted from HDR scenes captured under complex real-world illumination transitions. To this end, we initially introduce a tone mapping function which can be locally modulated by varying spatially (pixel-wise) its parameters as a function of the HDR content characteristics. Afterwards, we learn the local TMO parameters based on some characteristic features [1] extracted from the HDR image, using a support vector regressor (SVR) [175]. An essential problem in any learning-based approach is to find the right data to train the model. In the absence of an annotated dataset with optimal TMO parameters per pixel, we compute the ground-truth target parameters based on a set of HDR scenes captured under drastic illumination variations. For these scenes, we obtain per pixel ground-truth TMO parameters by solving an optimization problem, which simultaneously ensures: 1) stable keypoint detection; and 2) keypoint description robust to illumination changes. Since these two objectives are, in general, non differentiable, we also propose a proxy cost function which enables to compute the required derivatives and obtain an optimal solution.

The details of this work are available in the original paper [156].

3.2.1 Design of OpTMO

Let φ be a tone mapping function which maps the linear-valued HDR content of an image I to an output LDR I'. In general, for each image pixel x, the TMO operates as:

$$I'(x) = \varphi(I(x), \theta), \tag{3.2.1}$$

where $I(x) \in \Re$, $I'(x) \in [0, 255]$. The θ represents a set of parameters, given as $\theta = \{\theta_1, \theta_2, .., \theta_h\}$ where h is the number of parameters and depends on the considered TMO. Conventionally, these parameters are often tuned globally by cumbersome trial-test procedures to produce visually pleasing output images. Instead, we optimize the parameter vector θ per pixel, using support vector regression, in order to (1) to distinguish and localize a keypoint from its neighborhood locations; (2) to preserve local gradient orientation patterns around the keypoint; and (3) to bring invariance (as much as possible) to non-affine lighting variations in physical world scenes. To this end, we need to generate a training set of local parameters θ to use as ground-truth training data for the SVR, as explained in the following.

Ground-truth parameter generation

In order to find the ground-truth parameters θ, we employ a set of HDR images of the same scene captured under different lighting conditions, and search for the TMO parameters θ that maximize image matching performance after tone mapping *across all lighting conditions*. This provides a robust estimation of TMO parameters to use later in SVR training.

More specifically, we minimize a cost function $f(\theta)$ which drives the choice of parameters towards maximizing both keypoint repeatability and discriminability (see Section 3.1.2 and

[1] Here we use the term *characteristic feature* to denote any statistic computed from data, in contrast with the local features consisting of keypoints and descriptors.

3.1.3 for the definition of these measures). That is, our cost function is composed of two terms:

$$\underset{\theta}{\text{minimize}} \quad f(\theta) = E_{det}(\theta) + E_{des}(\theta), \tag{3.2.2}$$

where each term is computed over a scene composed of N HDR images with lighting variations. $P = \{(1,2),(2,3),\dots\}$ is the set of $K = \binom{N}{2}$ pair combinations of N images. The E_{det} term aims to ensure the co-variance of the corner response maps. Conversely, the E_{des} term helps in retaining the invariance of the discriminative patterns around the *key* locations in the image pairs when undergoing drastic transformations.

In order to maximize keypoint repeatability across pairs of images, we observe that it is important to enforce the similarity in detection response maps. We define the detection similarity term E_{det}, by summing the penalty computed from each pair in the set K, as:

$$E_{det} = \frac{\lambda_{det}}{K} \sum_{\{i,j\} \in P} \mathcal{C}_1(\mathcal{R}_i(\theta), \mathcal{R}_j(\theta)). \tag{3.2.3}$$

For each sample pair $\{i,j\} \in P$, we penalize the response maps dissimilarity by a logistic cost function given as:

$$\mathcal{C}_1(i,j) = \log(1 + \exp(\varepsilon_c - \langle \mathcal{R}_i \cdot \mathcal{R}_j \rangle), \tag{3.2.4}$$

where ε_c is the penalty control factor, $\mathcal{R}_i$ and $\mathcal{R}_j$ are the response maps corresponding to the images $i, j \in S$, and $\langle \cdot \rangle$ denotes the scalar product. λ_{det} weighs the penalization corresponding to detection. We employ the Harris cornerness measure [45] as response map $\mathcal{R}$, for its simplicity and effectiveness — we show experimentally that this choice still provides state-of-the-art repeatability even if a blob detector is used in practice.

The second term, E_{des}, aims to penalize the dissimilarity of the descriptors extracted from the tone-mapped images, and is defined as follows:

$$E_{des} = \frac{\lambda_{des}}{K} \sum_{\{i,j\} \in P} \mathcal{C}_2(\mathcal{D}_i(\theta) - \mathcal{D}_j(\theta)), \tag{3.2.5}$$

where $\mathcal{C}_2$ is the Euclidean distance and λ_{des} is a weighting factor. Since here we consider local (sparse) features, and not dense feature maps as in [155], we propose to constrain the penalization to those descriptors that belong to some potential keypoint region. That is, we compute the descriptor $\mathcal{D}$ after the keypoint localization which is obtained by applying the softargmax operation $\mathcal{S}$ on the resulting response map:

$$\mathcal{S} = \sum_i \frac{\exp(\beta z_i)}{\sum_j \exp(\beta z_j)} \cdot i \tag{3.2.6}$$

where z_i is the pixel location and β is a hyper-parameter for defining the shape parameter. The softargmax operation is a differentiable function to obtain local optima and helps in avoiding the cluttering in response maps. To compute an accurate keypoint localization, we define the

final gradient orientation around each pixel location computation as follows:

$$\mathcal{D} = \begin{cases} h(p), & \text{if } \mathcal{S}(\mathcal{R}) \geq \Lambda \\ 0, & \text{otherwise} \end{cases} \tag{3.2.7}$$

where $h(p)$ is the SIFT gradient orientation feature map over a patch p (see [156] for details), and Λ is the maximum softargmax value in a 16×16 neighborhood window of the considered pixel. It simply means that if the softargmax response score for the considered pixel location is maximum in its neighborhood window, only then the gradient orientation map is taken into account to contribute in the final descriptor-based penalty term in Eq. (3.2.5).

In order to find the ground-truth parameters θ We optimize the objective function in Eq. (3.2.2) using stochastic gradient descent (SGD).

Choice of tone mapping function

Many tone mapping approaches aim at separating scene illumination, which can display large dynamic range variations, from the reflectance of objects, which instead has lower dynamic range characteristics [22, 153]. Following this idea, we consider a tone mapping function φ, expressed as: $\varphi = I \cdot L^{-1}$. The illumination component L is estimated by an adaptive version of bilateral filtering [185] and is given as:

$$L(x,\theta) = \frac{1}{W} \cdot \sum_{y \in \Omega} \mathcal{G}_{\theta_1(x)}(\|x-y\|) \cdot \mathcal{G}_{\theta_2(x)}(\|I(x)-I(y)\|)I(y), \tag{3.2.8}$$

where $\mathcal{G}$ is a Gaussian kernel. The parameter map vector θ has two components, θ_1 and θ_2, also known as spatial and range variance. For each pixel location x, y is a pixel in the neighborhood Ω of x.

Training of support vector regression

Once the ground-truth target parameters have been computed, we train an SVR to predict θ based on some local characteristic features extracted directly from the HDR input image. Consider the sample set of characteristic features $\mathcal{F} = \{f_1, \ldots, f_n\}$ and the corresponding output denoted by $\mathcal{Y} = \{\theta_{k(1)}, \ldots, \theta_{k(n)}\}$ where $k = 1,2$ in our case. To build our predictor model, we want feature samples which capture distinctive information for both descriptor and detector. To that end, we build our feature sample f_i by concatenating two parts: a) the gradient-based SIFT pattern [104], 64 dimensional feature; and b) the 5×5 grid-based detector response feature [154], 25 dimensional feature. This forms a total dimension of 79. The features f_k are computed from the original HDR linear values, without any processing. This is not contradictory with the need to perform a TMO as, locally, HDR images generally display limited dynamic range [16].

3.2.2 Experimental results

We test our proposed OpTMO for image matching task on 8 HDR scenes at detection and description levels, and compare with state-of-the-art TMOs. The HDR dataset is composed of a total of 52 images, yielding a total of 280 test image pairs. We compare the proposed OpTMO with classical perception-based TMOs, including: the bilateral TMO (BTMO) [31], Chiu [22], Drago [29], Reinhard [159] and Mantiuk [112]. In addition, we also compare the OpTMO with optimizing *independently* either the E_{det} or the E_{des} terms in Eq. (3.2.2), respectively. Details about this independent optimization are available in [154] and [155]. We refer to these two methods as DetTMO and DesTMO, respectively.

We evaluate the results in terms of *Repeatability Rate* (defined in Section 3.1.2), *Matching Score* (defined in Section 3.1.3), and *Mean Average Precision* (mAP). The latter gives an overall evaluation of image matching, and is obtained by generating a precision-recall curve by varying the matching threshold. Recall is defined as the fraction of true positives over total correspondences and precision is given as the ratio of true positives to the total number of matches. Once the precision-recall curves are generated for each scene, we then compute the mAP scores by determining the area under the curves.

Keypoint detection

In Figure 3.2.1, we show the performance of our OpTMO and other state-of-the-art TMOs in terms of RR averaged over all test scenes, using several state-of-the-art keypoint extraction methods. For the sake of completeness, we also report the average RR obtained using HDR linear photometric values (HDRLin), without any tone mapping. Our results clearly show that the proposed OpTMO outperforms all the perception-based TMOs. In addition, the significant drop in performance with HDRlin demonstrates that HDR linear values are highly sub-optimal for keypoint detection task, similar to what is found in previous studies An analysis of per scene performance shows that gains are especially important for indoor scenes, which have been acquired by varying locally the illumination and introducing stark shadows. Notice that Harris detector here tends to be favored by our method, as we use the Harris cornerness response $\mathcal{R}$ in Eq. (3.2.4); however, even totally different detectors, such as SIFT (blob) detectors, get significant improvements in repeatability using the proposed TMO. We also observe that keypoint repeatability for OpTMO is lower compared to DetTMO. This is expected, as the additional descriptor-level cost term in Eq. (3.2.5) changes the objective function with respect to detector repeatability only (as in DetTMO). However, the joint optimization enables to increase the overall matching performance, as showed in the following.

Descriptor matching

In Figure 3.2.2, we compare the average OpTMO matching score with respect to state-of-the-art TMOs, for several descriptors (gradient-based or binary). Overall, we attain significant gains in terms of MS using all feature extraction methods. The gains are more important for gradient-based features schemes such as SIFT and SURF, which is expected by design given

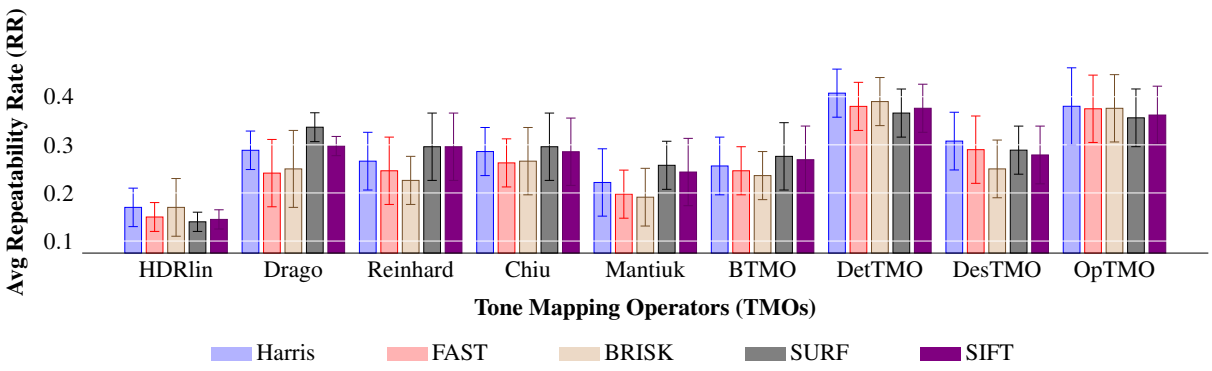


Figure 3.2.1: **Keypoint Detection**: Average Repeatability Rate computed on different TMOs using various keypoint detection schemes. The average is calculated over all test scenes.

the definition of the descriptor signature in Eq. (3.2.7).

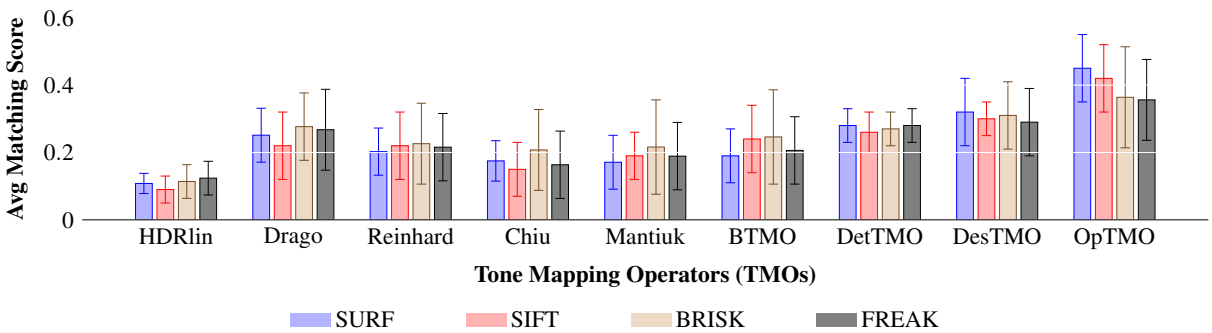


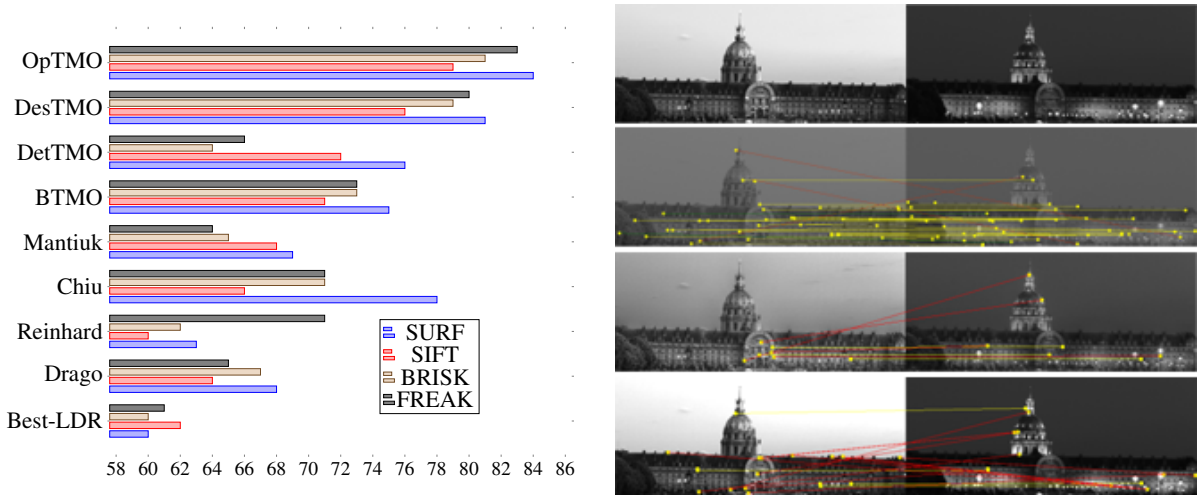
Figure 3.2.2: **Descriptor Matching** computed on different TMOs using SURF, SIFT, FREAK, BRISK descriptor extraction schemes. The average is calculated over all test scenes.

Image matching

We evaluate the full image matching chain by computing mean average precision (mAP) scores over the complete dataset. The results per TMO are reported in Figure 3.3(a). We observe that for every descriptor extraction scheme our proposed model outperforms all the other TMOs. High mAP scores imply that our model obtains more correct matches and reduce the probability of false matches. An illustration of matching results is given in Figure 3.3(b), showing that the proposed full-chain optimal tone mapping improves the matching efficiency in drastic lighting variations. Notice that ReinhardTMO and MantiukTMO provide poor image matching results compared with the proposed approach, although they provide better visually looking images.

3.3 Detection of inverse tone mapping in HDR images

Image forensics is a well recognized research field in multimedia security, which aims at achieving image authentication in a blind and passive manner [148]. Various image forensic methods



(a) mAP % scores (b) Day/Night matching using SURF.

Figure 3.2.3: **Image Matching**. (a) mAP % scores for the 9 different LDR modalities using 4 feature extraction schemes. Scores are averaged over 8 scenes. (b) An example of matching: Row I: 2 HDR luminance images from *Invalides*, scene are displayed after log scaling. Row II: the feature matching using our proposed OpTMO (21 correct and 3 incorrect matches). Correct and incorrect matches are shown with yellow and red lines, respectively. Row III: using Mantiuk TMO (3 correct and 4 incorrect matches). Row VI: using Reinhard TMO (3 correct and 11 incorrect matches).

have been proposed in the literature; however, all of them have been designed with the conventional, 8-bit low dynamic range (LDR) image representation in mind. In this section, I consider instead a forensic problem related to the *source identification of high dynamic range images*.

Nowadays, the two most common techniques to generate HDR content include: 1) acquiring multiple conventional LDR pictures of the scene at different exposure times, which can be fused together afterwards using, e.g., the method in [27] — we will refer to HDR pictures generated in this way as **mHDR**; 2) acquiring an LDR picture of the scene, and expanding its dynamic range through an operation commonly known as *inverse tone mapping* (iTM), since conceptually it does the opposite of tone mapping algorithms conceived to display HDR pictures on LDR displays [7]. We will refer to this kind of images as **iHDR**. This latter option is particularly attractive considering that nowadays the majority of legacy video footage is LDR, and that range expansion is needed to display it on next-generation HDR displays [170]. Furthermore, it has been shown that in many cases HDR video obtained through iTM yields similar, or even indistinguishable, visual experience as HDR content generated by multiple exposures [2, 26]. In this work, we consider the forensic problem of **identifying whether an HDR picture is mHDR or iHDR**. From a multimedia security perspective, solving this forensic problem can help to identify the authenticity of a content, and to localize tampering whenever mHDR and iHDR image patches have been composed together to create a forgery. To this end, we propose a forensic feature able to perform fine-grained mHDR/iHDR classification, in such a way to accurately localize the tampered regions and infer their semantics consequently.

The details of this work are described further in the original papers [34, 35].

3.3.1 Forensic analysis based on Fisher scores

HDR image forensics presents some subtleties and new challenges with respect to standard forensic techniques. For instance, while iTM might resemble a contrast enhancement process, classical forensic detectors based on statistical fingerprints [176, 20] fail when applied on iHDR pictures, as those images do not present typical peak/gap artifacts in their histogram. In [35], we analyzed the performance of a Fourier-based forensic feature that describes the presence of periodic patterns in iHDR pictures, based on the observation that iHDR pictures are obtained from an LDR signal, which has a discrete nature. We found there that such a detector achieves about 85% accuracy when iHDR images are obtained from 8-bit LDR pictures. However, performance rapidly decrease when one considers iHDR obtained by 16-bit RAW images output by professional digital cameras. This shows that second-order statistics are only partially effective. In fact, as a result of the continuous nature of HDR values, most forensic methods based on integer arithmetics, such as classical compression and point-wise transformation detectors [36, 10, 176] are not applicable.

More sophisticated forensic tools employ higher-order statistics to model local image content, e.g., the Subtractive Pixel Adjacency Matrix (SPAM) features [144]. These methods compare neighboring pixel values as if they were lying on a uniform interval scale. However, as see in Section 2.1, HDR images need to be preprocessed through a non-linear transformation in order to yield perceptually uniform values. Our proposed forensic feature also employs higher-order statistics, which are modeled by means of Gaussian mixture models (GMM).

Fisher scores

In the LDR image analysis literature, the concept of Fisher scores [63] has largely influenced image classification in the form of the well-known Fisher vector [143], which further inspired the recently proposed LHS feature [171] in texture/facial analysis. Here, we use this concept for the first time in a HDR forensic scenario.

Given a generic pixel $\mathbf{z}_0$ and its $s \times s$ local neighborhood $\{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_{s^2-1}\}$, we obtain a local differential vector $\mathbf{x}$ with $\mathbf{x}_i = \mathbf{z}_i - \mathbf{z}_0$ $(i = 1, 2, \cdots, s^2 - 1)$. Its likelihood under an M-component Gaussian mixture model (GMM) parametrized by $\theta = \{\pi^i, \mu^i, (\mathbf{C}^i) | i = 1, 2, \cdots, M\}$ is computed as:

$$L(\theta|\mathbf{x}) = p(\mathbf{x}|\theta) = \sum_{i=1}^{M} \pi^i \mathcal{N}(\mathbf{x}|\mu^i, \mathbf{C}^i), \tag{3.3.1}$$

where π^i, μ^i, and $\mathbf{C}^i$ are respectively the mixing weight, mean, and covariance matrix of the i-th GMM component.The higher-order statistics in the local neighborhood of $\mathbf{z}_0$ can therefore be represented using the *Fisher scores* [63], which are calculated as the partial derivatives with respect to the parameters θ of the log-likelihood, *i.e.*:

$$\mathbf{F}(\theta, \mathbf{x}) = \nabla_\theta \log L(\theta|\mathbf{x}). \tag{3.3.2}$$

For the sake of simplicity and also for reducing the dimensionality of the final forensic feature, we assume $\mathbf{C}^i$ are diagonal matrices. Since $\mathbf{x}$ captures the local derivatives of $\mathbf{z}$, which carry

high-frequency information, the trained means μ^i are generally close to zero. Therefore, we remove the DC component before training the GMM models. The resulting Fisher score vector $\mathbf{F}(\theta, \mathbf{x})$ has a length of s^2M.

In practice, we compute the Fisher scores with respect to *two* GMMs, parametrized by θ^0 and θ^1, representing mHDR and iHDR, respectively. These two GMMs are learned in a previous, off-line training stage from a database containing the two classes of HDR images, using the Expectation-Maximization algorithm. Given $\mathbf{x}$, we then can form the following $2s^2M \times 1$ sized Fisher score vector:

$$\tilde{\mathbf{f}}(\mathbf{x}) = \left[\mathbf{F}(\theta^0, \mathbf{x})^T, \mathbf{F}(\theta^1, \mathbf{x})^T\right]^T, \tag{3.3.3}$$

which is further normalized to construct the proposed forensic feature vector with the i-th element as:

$$\mathbf{f}_i(\mathbf{x}) = \operatorname{sign}\left(\tilde{\mathbf{f}}_i(\mathbf{x})\right) \frac{\left|\tilde{\mathbf{f}}_i(\mathbf{x})\right|^{1/2}}{\sum_i \left|\tilde{\mathbf{f}}_i(\mathbf{x})\right|}, \; i = 1, 2, \ldots, 2s^2M. \tag{3.3.4}$$

After defining the forensic feature in Eq. (3.3.3), we use it in a *discriminative* model such as a Support Vector Machine (SVM) in order to classify images as mHDR or iHDR.

3.3.2 Experimental results

In order to conduct our forensic tests, we collected a dataset of 498 high-resolution mHDR images from several publicly available HDR datasets, as detailed in [35]. For iHDR, we consider an equal number of LDR pictures, and we expand their dynamic range using 6 popular iTM algorithms: Akyüz *et al.* [2] (**A**); Banterle *et al.* [8] (**B**); Huo *et al.* [53] (**H**); Kovaleski and Oliveira [86] (**K**); Meylan *et al.* [116] (**M**); Rempel *et al.* [160] (**R**). **A** performs a linear expansion of the dynamic range to match a target HDR display range; **B**, **H**, **K** and **R** expand the range using a nonlinear function, and adjust the dynamic range locally by means of an expand map; finally **M** applies different linear expansions in different areas of the image, which are classified as diffuse or specular.

In order to train the SVM classifier to detect iHDR/mHDR, we crop 512×512 images from the pictures in the datasets mentioned above. This gives a total of 1839 iHDR/mHDR images for each of the six iTM for training, and 1851 iHDR/mHDR images for testing, respectively. We also consider smaller block sizes, up to 3×3 pixels. Finally, we also consider the '**mix**' scenario to train and test forensic features, where all the iTM algorithms are mixed together. Details about the construction of train/test datasets are reported in [35]. We compare the proposed feature with 2nd-order SPAM features [144], and with the LHS feature [171], which is used in face recognition and that is similar in principle to the proposed one, although it uses a single non-zero-mean GMM to compute Fisher scores. 38-component and 43-component GMMs are used respectively by the proposed method and the LHS feature. For all the considered methods, we compress the HDR pixel values by applying a logarithmic function prior to feature extraction.

Detection accuracy for different methods, iTM and image size are reported in Table 3.3.1. The proposed method achieves at least comparable performance with the SPAM/LHS features, and

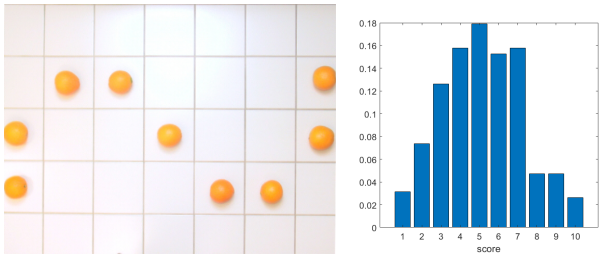
Table 3.3.1: Detection accuracy (%) comparison when different image (block) sizes are considered. The feature dimensionalities of the SPAM feature, the LHS feature, and the proposed feature are respectively 686, 688, and 684.

size	feature	**'mix'**	**'A'**	**'B'**	**'H'**	**'K'**	**'M'**	**'R'**
512×512	SPAM	**97.19**	**97.16**	96.92	**97.08**	**97.14**	**97.00**	**97.92**
	LHS	94.68	94.79	96.38	95.14	94.95	95.11	96.81
	Proposed	94.35	93.84	**97.11**	94.06	94.38	94.81	97.27
8×8	SPAM	73.56	74.37	71.01	73.75	74.15	74.31	79.76
	LHS	72.98	73.75	69.69	71.49	73.52	73.76	77.88
	Proposed	**76.45**	**76.70**	**75.55**	**76.21**	**76.59**	**76.37**	**81.24**
7×7	SPAM	71.39	72.25	68.07	71.26	72.37	72.36	78.36
	LHS	70.12	71.66	68.26	70.57	71.11	71.40	76.48
	Proposed	**74.67**	**74.99**	**73.95**	**74.20**	**74.79**	**74.86**	**80.32**
6×6	SPAM	69.68	70.96	66.35	70.21	70.43	70.93	75.88
	LHS	69.12	70.54	66.08	69.69	70.61	70.15	75.63
	Proposed	**72.70**	**73.33**	**71.87**	**72.01**	**73.01**	**73.12**	**78.22**
5×5	SPAM	67.10	68.25	63.24	67.26	68.17	68.50	74.23
	LHS	67.72	68.21	64.05	67.09	68.81	67.88	72.81
	Proposed	**70.82**	**71.16**	**68.64**	**70.68**	**71.38**	**71.36**	**76.52**
4×4	SPAM	63.06	63.97	59.22	63.11	63.87	64.41	69.76
	LHS	64.67	65.09	60.32	64.49	65.19	65.60	69.98
	Proposed	**67.50**	**67.71**	**64.67**	**67.25**	**67.55**	**67.91**	**73.88**
3×3	SPAM	-	-	-	-	-	-	-
	LHS	62.28	64.02	58.66	62.59	63.82	63.76	68.93
	Proposed	**63.66**	**64.89**	**60.44**	**63.92**	**64.79**	**64.63**	**70.51**

is especially advantageous on *very small* image blocks. Note that, on image blocks as small as 3×3, the SPAM feature cannot even be extracted, as it is not possible to count the co-occurrences of neighboring second-order derivatives. However, the proposed method can still perform the forensic task thanks to the fact that the GMMs are learned on 3×3 image patches. Though in such an extreme case the detection accuracies are much lower than for 512×512 images, we believe that these result show the boundary achievable by forensic methods when we keep pushing the limits of image block size. This is very important for the forensic study of *very fine-grained* image tampering localization.

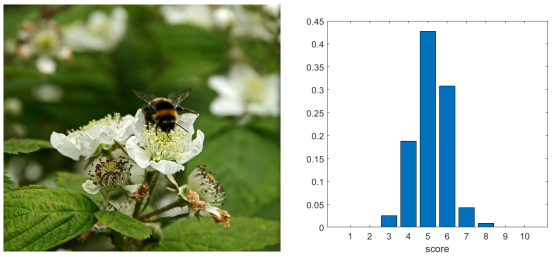
3.4 Predicting subjectivity in image aesthetic assessment

I conclude this chapter with recent work on predicting the aesthetic value of a picture. The goal of image aesthetic quality assessment is to determine how beautiful an image looks to a human observer. This problem lies in between quality assessment, as it aims at quantifying the visual experience of watching a photo, and image analysis, since aesthetics is definitely affected by higher-level features such as image content and other semantic clues.



Mean	SD	MAD	MED	DUD
5.174	2.104	1.697	0.278	3.570

(a)



Mean	SD	MAD	MED	DUD
5.180	0.902	0.657	0.650	4.439

(b)

Figure 3.4.1: Example of aesthetic subjectivity for two images of the AVA dataset. The two images, displayed in the top-left panels, have similar mean score but different distribution of aesthetic judgments given by human raters, shown in the histograms on the top-right panels. The tables report several measures that compactly describe subjectivity based on the score distribution, which are described in Section 3.4.1.

More precisely, I deal here with the intrinsic *subjectivity* of aesthetic assessment, which derives from the different opinions human raters might have about the beauty of a digital picture. Most of existing aesthetic quality prediction approaches assume that aesthetic quality can be represented by a single value, e.g., the mean aesthetic score or the aesthetic class (good/bad). However, this assumption does not take into account other internal factors that may influence the aesthetic judgments, such as personal background, interests, mood, etc. Indeed, experimental psychology studies show that, while beauty is conveyed by objective visual clues, the resulting aesthetic appraisal is subjective and depends on how the visual clues are processed by higher-level cognitive areas in the brain.

In the rest of this section I discuss our preliminary work on quantifying and predicting subjectivity in image aesthetics. Specifically, we first introduce scalar measures to quantify subjectivity based on the distribution of aesthetic scores; afterwards, we compare two prediction schemes to estimate these measures. This work is described in greater details in [67].

3.4.1 Subjectivity measures

We define subjectivity as the *degree of consensus* about the aesthetic value of a picture when the latter is judged by a panel of human raters. The top row of Figure 3.4.1 illustrates this with an example: two images from the AVA dataset (which is one of the largest available subjectively annotated aesthetic dataset) have a similar average aesthetic score, but a different degree of subjectivity. In the image in Figure 3.1(a), it is evident that humans tend to agree more on the aesthetic quality of the image, while the judgments are more dispersed for the image in Figure 3.1(b). Intuitively, being able to predict aesthetic subjectivity can provide valuable information in order to determine to which extent aesthetic predictions can be trusted. This in turn could be beneficial in applications such as enhancement or retrieval, in order to obtain more reliable and accurate results.

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

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

- **Standard Deviation** (SD) of the score distribution, which describes the dispersion of the scores around the average score. A higher value of SD indicates a lower consensus around the average score, and thus higher subjectivity.

- **Mean Absolute Deviation around the median** (MAD), defined as the sample average deviation of the scores around the median score, that is:

$$MAD_n = \frac{1}{M_n} \sum_{i=1}^{M_n} |x_n(i) - m_n|. \tag{3.4.1}$$

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

- **Distance to Uniform Distribution** (DUD). We consider the distance of the score distribution $\mathbf{p}_n(x_n)$ from the distribution having the maximum entropy over s, which is the uniform distribution. We quantify this distance using the 2-Wasserstein metric $d_W(\mathbf{p}_n, \mathbf{u}_s)$, that is:

$$DUD_n = d_W(\mathbf{p}_n, \mathbf{u}_s) = \left[\sum_{i=1}^{k} (\mathbf{P}_n(i) - \mathbf{U}_s(i))^2 \right]^{1/2}, \tag{3.4.2}$$

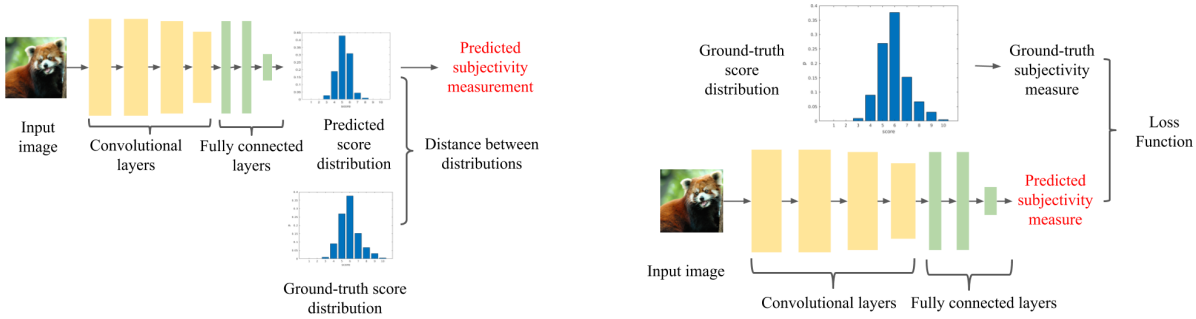
  where $\mathbf{u}_s$ is the discrete uniform distribution defined over the categories s, and $\mathbf{P}_n$ and $\mathbf{U}_s$ are the cumulative distribution functions of $\mathbf{p}_n$ and $\mathbf{u}_s$, respectively.

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

- **Distance from the Maximum Entropy Distribution** (MED). Since the uniform distribution has always a mean value equal to the midpoint of the score scale, the DUD measure tends to penalize more skewed distributions having mean values close to the extremes of the quality scale. To overcome this bias, we compare the score distribution with the maximum entropy distribution over the quality scale having the *same mean*. More specifically, we look for a discrete distribution $\mathbf{q}_s$ which solves the following optimization problem:

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

  where H denotes discrete entropy and $\mu[\mathbf{q}]$ is the mean of $\mathbf{q}$.



(a) Indirect subjectivity prediction

(b) Direct subjectivity prediction

Figure 3.4.2: Subjectivity Prediction Framework. In the indirect prediction framework, an aesthetic score distribution is estimated first, and subjectivity measures are computed over it. We compare this approach with directly predicting subjectivity computed on ground-truth distributions (b).

Then MED for image n is defined as:

$$MED_n = d_W(\mathbf{p}_n, \mathbf{q}_s) = \left[\sum_{i=1}^{k} (\mathbf{P}_n(i) - \mathbf{Q}_s(i))^2 \right]^{1/2}, \tag{3.4.3}$$

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

The tables in Figure 3.4.1 show an example of these measures computed for the two images in the top panel. We can observe that all of them capture correctly the degree of consensus of the score distributions. In the following, we will study how accurately each of these measures can be predicted, either directly or by means of predicted score distributions.

3.4.2 Prediction of subjectivity

In order to predict the subjectivity measures proposed above, we consider two options: i) we predict the score distribution *indirectly* using an existing score prediction method [65, 66, 127]; or ii) we compute subjectivity measures on ground-truth scores, and learn to predict them *directly*. These two approaches are illustrated in Figure 3.2(b), and compared experimentally in the following.

Experimental setup

We predict subjectivity using a deep convolutional neural network. We employ a modified Resnet-34 network as baseline for our experiments. In addition, we also use Resnet-101 to study the influence of a deeper network structure in the direct aesthetic subjectivity prediction. The details about the network architecture and training are available in [67].

Table 3.4.1: Pearson's Linear Correlation Coefficient (PLCC)

Methods	***SD***	***MAD***	***MED***	***DUD***
Bin Jin's[65]	0.145	0.159	0.178	0.096
NIMA [181]	0.169	0.187	0.211	0.255
RSCJS [66]	0.187	0.199	0.227	0.281
Direct (Resnet-34)	0.274	0.276	0.323	0.351
Direct (Resnet-101)	**0.307**	**0.304**	**0.333**	**0.360**

We use the standard training/testing partition of the AVA dataset [128] to learn directly or indirectly subjectivity. The AVA dataset contains over 250,000 images from photography amateurs. The scores have been collected through approximately 1400 photographic challenges from viewers who voted integer scores in the range $[1, 10]$. The number of votes in AVA ranges between 78 and 549, and the average is around 210, thus enabling a more reliable estimation of score distributions.

For the indirect subjectivity prediction, we consider the following three methods for predicting aesthetic score distributions: the work of Bin Jin et al. [65] (chi-square distance loss); NIMA [181] (Earth Mover's Distance loss); and the RSCJS method of Jin et al. [66] (cumulative Jensen-Shannon divergence loss).

Experimental results and discussion

Table 3.4.1 reports Pearson's linear correlation coefficients (PLCC) between the predicted subjectivity measures, and the subjectivity measures computed on ground-truth aesthetic scores. We observe that direct subjectivity prediction always outperforms indirect prediction through distribution scores, for all the proposed subjectivity measures. In particular, for the same network complexity (Resnet-34), predicting directly the subjectivity is clearly better than predicting the score distribution first and computing subjectivity based on it. Similar results are obtained when considering Spearman correlation coefficients and RMSE, as reported in [67].

Although direct prediction improves all the considered performance indicators, we observe that overall the prediction performance is still not satisfactory, e.g., the PLCC is below 0.4. We might wonder whether this is due to a limited capacity of the Resnet-34 model we employed. Therefore, in order to study how subjectivity prediction performance improves with a more complex network, we tested the direct prediction scheme using Resnet-101, which is much deeper than Resnet-34. As expected, the results generally improve over the simpler Resnet-34. However, this improvement is in most case only marginal, showing that aesthetic subjectivity prediction is intrinsically a hard problem – at least a harder one than predicting the average aesthetic score, where PLCC between predicted and ground-truth values is higher than 0.6 [181].

Comparing the different subjectivity measures, those inspired by information theory (DUD and MED) are in general those with higher prediction performance. Among the statistical motivated descriptors, the SD is generally predicted more accurately than MAD. We can assume that, for the same neural network model complexity, a ground-truth variable which has a higher dependence on the input is easier to predict, or, in other terms, target variables which tend to be

more "noisy" will be more difficult to learn. Thus, we can argue that the subjectivity measures based on information theory are somewhat more robust than statistical deviation measures. A possible rationale behind this could be that both DUD and MED are based on distances between histograms, which take into account the whole score distribution. On the other hand, SD completely captures data variability when the underlying score distribution is Gaussian, which is the case for only 62% of AVA images [66]. MAD is supposed to be more robust to skewed distributions, but it might be affected by the sample median computation, which on a 10-dimensional distribution as for aesthetic scores can only take values over a small set, i.e., $\{1, 1.5, 2, \ldots, 10\}$.

Despite our approach achieves state-of-the-art subjectivity prediction performance, predicting subjectivity is still a very challenging task. We believe that this is partially due, in addition to the complexity of the task in itself, to the noisy nature of current aesthetic datasets. This is evident for the benchmark AVA dataset, where aesthetic scores are influenced by many factors that go beyond the pure aesthetic value of a picture. Building cleaner and more reliable aesthetic datasets is among the future directions to consider in this field.

Part III

Future research perspectives

Chapter 4

Towards effective representations for immersive visual communication

Finding meaningful and effective signal representations is one of the most fundamental problems in signal processing. In compression, modeling some properties of a signal, such as piecewise smoothness and other forms of regularities, enables to efficiently encode and transmit data. In visual quality assessment, simple pixel-wise metrics such the mean squared error are known to be unable to capture perceptual and contextual phenomena that incur in the formation of quality judgments. Instead, computing distances using more advanced visual representations can display better predictions of human opinion scores. In general, visual analysis tasks substantially rely on designing effective representations and features able to capture the properties of interest in the signal.

For human perceptual tasks, the cognitive process through which visual information is analyzed requires a complex and accurate modeling of the human visual system. This is especially true for immersive video formats, where the goal is to produce a truly realistic and interactive user experience, which involves, e.g., 3D and free viewpoint perception. In the past few years, the problem of designing good features and representations has been partially solved by *data-driven* techniques, thanks to tools such as deep neural networks, and to the availability of large annotated datasets. The success of this approach is evident in tasks which traditionally are difficult to model, such as classification or detection, where machine learning is able nowadays to achieve better performance than an average human annotator. Nevertheless, in domains such as visual communication, tools such as deep representation learning and generative models have become popular only recently (see Section 1.3 and 2.2), and their applications to visual communication is an emerging research topic. In image coding, for instance, end-to-end schemes based on deep representation learning, such as auto-encoders, are still unable to provide competitive results compared to state-of-the-art hybrid codecs [191]. Learning representations end-to-end in the case of video is fundamentally limited by computational complexity and memory issues, and will be an important research topic for the next years. In quality assessment, learning what is a "natural" (undistorted) image is significantly more difficult than modeling single-class image manifolds (such as faces or specific objects), due to the very large variability of the content. In this context, learning representations able to generalize to real-world content variations remains an open problem.

In order to conceive effective representations for immersive video, assessing the quality of experience (QoE) is of fundamental importance. Emerging 3D representations such as point clouds, for example, entail new kinds of processing artefacts, such as geometric distortion. How to properly assess these artefacts, whose perception is the result of interaction between the observer and the content, is currently an active research area. More generally, data collection is of paramount importance in data-driven approaches, and even more so in QoE, where ground-truth scores can not be objectively measured as in other computer vision applications such as tracking or classification, but are subjective and stochastic in nature. A typical example is predicting the aesthetic value of a picture, as discussed in Section 3.4. One of the greatest challenges in data collection for QoE is the trade-off between the quantity and the quality of subjective data. In the example of aesthetics, the largest datasets used nowadays have been collected "in the wild": images are downloaded from existing online resources, where aesthetic labels are often confused with other concepts such as popularity, interestingness, etc. This fundamentally limits the possibilities to extract deeper knowledge from data, e.g., which are the factors that make a picture look beautiful. Designing proper methodologies to collect data in a disciplined way, following QoE best practices, under uncontrolled conditions, is a key factor to enable further advances in the understanding of complex perceptual attributes of images.

My research project is at the intersection of three main axes, as illustrated in Figure 4.1.1: i) designing good visual representations for compression and quality assessment; ii) conceiving new methodologies for measuring QoE and collecting subjective data; iii) applying these tools to achieve higher immersion in visual communication. In the following, I will detail these axes, pointing to mutual interactions across them, and outlining some possible research directions in each domain.

4.1 Learning good representations for video compression and quality assessment

Deep generative models — from restricted Boltzman machines to variational auto-encoders and, more recently, generative adversarial networks [42] — can learn effective representations for very complex signals such as natural images. I will consider how to use these representations both in compression of visual content and in visual quality assessment.

4.1.1 Deep generative models for video coding

In the past few years, deep generative models and representation learning have been shown to be able to effectively capture the complex statistics of image and video. Recently, these tools have been applied to image and video compression, displaying potential gains compared to traditional architectures. As a result, learning-based compression is nowadays a hot research topic, as demonstrated by the popularity of challenges such as CLIC-CVPR, and by the increasing interest towards these approaches in standardization committees such as JPEG (e.g., the new JPEG-AI initiative) and MPEG.

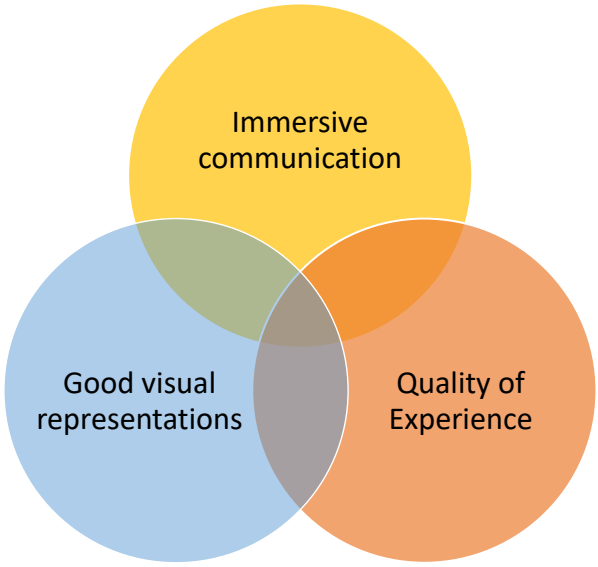


Figure 4.1.1

The goal of generative models is to learn the distribution of the data. This can be effectively achieved when the data variability is relatively small, e.g., when images are drawn from a few semantic classes. However, learning good generative models becomes significantly more difficult when the input data is unconstrained, which is instead the typical use case for video compression. Indeed, recent work on learning-based spatial prediction shows that classifying input patches into two classes according to their prediction characteristics, and learning a model for each of the classes, is in general better than learning a single, "one-size-fits-all" model [98]. However, the problem of establishing automatically how many and which are the classes that optimize the coding efficiency has not been studied so far. Differently from the conventional discriminative classification problem, the number of classes, as well as the frontiers between them, is not well defined in advance, nor it corresponds to an explicit semantic labeling of the training data. Instead, since the final goal is coding efficiency, the classification criterion is rather linked to the rate-distortion performance of the specialized codec which follows the learned classifiers. Therefore, an interesting direction will be to study how to jointly learn classification and coding.

Another application of deep generative models to video compression is in context-based *entropy coding*. Arithmetic coding enables to code symbols by approaching the entropy rate of the source, provided the data distribution is known. When only an approximation of the data distribution is available, there is a loss of coding efficiency, which is quantified by the Kullback-Leibler divergence between the actual distribution and the used one. The context-adaptive binary arithmetic coder used in modern video codecs employs a set of predefined conditional probability models, which are updated on-the-fly during encoding depending on the observed symbols. The problem of context dilution — the degradation of probability estimation accuracy due to the curse of dimensionality — limits the number of different contexts that could be used. We will consider a different approach: instead of using several conditional probabilities tables, which practically forces the context to be small, we will instead use a deep generative model

with a large context, which could potentially involve all the previous coded symbols. To this end, we will consider recurrent neural networks to model sequences of symbols; this approach has been successfully used, e.g., in image generation [44].

Finally, a fundamental challenge in video coding is modeling the operational *rate-distortion* (RD) function of the codec. This problem is know to be hard, for several reasons. First, state-of-the-art video codecs are very complex systems, which use several tools in addition to quantization in order to yield high compression efficiency. Some of these tools are difficult to model, e.g., mode decision. Second, the RD performance of a codec depend strongly on the content. Finally, modeling coding dependencies induced by predictive coding is a well-known chicken-and-egg problem [190], which would require architectures able to take into account the state of the prediction such as recurrent neural networks. Deep neural networks are universal approximators, and could be used in principle to learn a direct mapping from pixels (plus some side information) to a point in the RD plane. However, training this kind of system, especially on a fine-grained coding unit scale, is not trivial, as it somehow depends on the choice of the coding mode. Learning the RD function would be extremely useful in rate-distortion optimization, and especially to embed a bit-rate term in the loss function for learning-based coding tools as the context-based prediction enhancement described in Section 1.3. Moreover, approximating the RD function with a neural network would solve one of the greatest issues in end-to-end learning-based compression methods: dealing with non-differentiable operations such as quantization. This is currently approximated in various ways, e.g., by replacing quantization noise with uniform noise in backpropagation. Instead, a neural network-based rate estimator is by construction differentiable and would easily fit a gradient optimization technique.

4.1.2 Representation learning in quality assessment

As mentioned above, representing visual content in a proper space is key in evaluating visual quality, by incorporating the cognitive process of the human visual system. This cognitive process can be seen from two different perspectives. On one hand, quality assessment methods can be framed as the problem of finding a *discriminative* model for the data, i.e., by performing a supervised training of a proper regressor in order to explain and predict a set of observations. These consist of (distorted) images associated to their mean opinion scores. Deep convolutional neural networks have been successfully used to this purpose with impressive results, as shown in Section 2.2, both to predict overall quality scores and to localize the distorted regions. The main disadvantage of this approach is the need for large annotated datasets, which in the case of quality scores requires a costly and cumbersome data collection work.

On the other hand, one can cast visual quality assessment as the problem of finding a *generative* model for visual information. Estimating this model is an unsupervised learning problem, which has been addressed in different ways in the past. A large amount of work has been done to model natural scene statistics (NSS), e.g., in terms of the distribution of wavelet coefficients over large collections of undistorted images. Deviations with respect to these statistics have been used to quantify the impairment of an image [125]. However, evidence from neuroscience and cognitive theories (from seminal Helmholtz's studies in the 1860's and Gestalt school principles, to modern neuroscience tenets) suggest that vision is the outcome of an active inference

process and not simply a matter of signal representation. In this respect, psychovisual quality is closely related to how accurately visual sensory data can be explained by the brain internal generative model. This principle has been successfully employed in the recent "free-energy" based quality approach [212], where the generative model is approximated by a simple auto-regressive process. Clearly, this is only a first-order approximation, as the internal brain model is far from being fully understood. In this context, deep neural network architectures seem particularly promising, as their cascaded nonlinear transformations are believed to mimic the processes of evolution that have shaped visual representations within the human brain.

In this context, I am particularly interested in learning representations using an opinion- and distortion-unaware approach, such as in the NIQE quality metric [123]. While there simple local image representations are used, such as the mean-subtracted contrast-normalized pixels (see Section 2.2), it is reasonable to expect that higher level representations learned, e.g., by a deep auto-encoder, might lead to a more expressive quality representation. In addition, how to compute distances between these representations is itself an open question. Previous work has modeled coefficients in the feature space using generalized Gaussian distributions [123], which enables to easily compute the likelihood of a sample under this model, or to compute distances between distributions. In the general case, one could instead learn this distance function from data. In alternative, one could learn representations in such a way to maximize the distance between pristine and distorted images, using a self-supervised approach in which pristine images are corrupted with different kinds of artefacts, and using architectures such as Siamese networks to learn embeddings in the representation space.

4.2 New methodologies for Quality of Experience

As mentioned above, a fundamental challenge in learning representations for visual quality is the trade-off between the quantity and quality of subjective data. Collecting good data is crucial to learn good representations. However, data collection in QoE is costly and intrinsically subjective, it can require specialized equipment and needs a precise definition of the task to evaluate. An important aspect in collecting QoE at large scale is using efficient methodologies to sample the distribution of opinion scores. A promising research area is employing *active sampling* approaches [100], and fusing datasets collected using different methodologies as discussed in Section 2.4.

There has been a large amount of work on defining recommendations for assessing visual QoE in multimedia services [57, 62], as well as studies on how to collect user opinions using crowd sourcing approaches. While this is feasible for simple quality tasks, more advanced and *multi-dimensional* QoE problems, such as judging the aesthetic value of pictures, require a careful design in order to collect meaningful data. Specifically, the multi-dimensional nature of complex QoE tasks requires defining clearly and unambiguously the perceptual/semantic dimension to collect, that should be explained to observers through a proper training phase as in well-established best practices of QoE assessment. This might include as well indicators about the confidence or uncertainty of observers in giving their judgments. A more solid understanding of these good practices in multimedia QoE and how to apply to large-scale data collection is an ongoing research field, which can largely benefit from studies in other related fields such as

psychology, sociology, marketing, etc.

Measuring QoE for *immersive* content is substantially different from QoE of conventional 2D video. In fact, standard evaluation protocols explicitly require that the viewing conditions be the same across observers. These include the viewing distance and angle, the lighting conditions, and, of course, the stimuli, which (up to a permutation) are rendered in the same way for each observer. While these conditions were appropriate for classical 2D television, in immersive video each observer can explore a scene in a different, personal way. As a result, a number of natural questions arise when assessing visual quality of experience of humans in an interactive scenario: How to adapt the existing subjective assessment methodologies to consider interaction? How does the interaction between the human and the visualization device affect the perceived quality and visual attention? Which is the most appropriate device to collect ground-truth subjective scores? These will be interesting research directions to explore for emerging immersive formats such as 3D point clouds and light fields.

4.3 Immersive visual communication

Immersive communication consists in producing a *realistic and credible* virtual communication experience in which remote users feel as they were exchanging the same stimuli and receiving the same feedback from the environment as if they were participating to a face-to-face meeting. Producing truly realistic and immersive 3D video experience is widely considered as the key for the success of applications such as virtual (VR) and augmented reality (AR). Immersive AR/VR products have started to appear in the consumer market, revolutionizing a broad range of sectors, from immersive communication (e.g., Microsoft Holoportation), to telemedicine, design, and even working habits. This has been made possible by the development of efficient immersive video representations beyond conventional 2D images and video. The most general of these representations, known as *plenoptic function*, aims at reproducing the intensity of light seen from any viewpoint or 3D spatial position, angular direction, over time and for each wavelength. In practice, this would require a huge amount of information to be captured and stored, and therefore, several different samplings (and thus, approximations) of it have been proposed.

One attractive way to approximate the plenoptic function from the point of view of human interaction is to shift from a flat, 2D pixel-based format to a geometric representation such as *3D point clouds*, which provide six degrees of freedom interaction with objects in the scene. Point cloud video representations come with a number of technical challenges. The non-regular sampling of point clouds makes difficult to use conventional signal processing tools, which have been traditionally designed to work on regular discrete spaces such as a pixel grid. This is particularly critical for point cloud video compression (PCC), which is essential to store and transmit the large amount of geometric information of a scene (see Section 1.4). Future research in immersive point cloud visual communication will definitely benefit from advances in representation learning and QoE assessment, as discussed above. For instance, current approaches for PCC use octree-based representations for geometry and attribute compression, or 2D projections in the case of dynamic PC. We have been among the first to propose to learn representations for voxelized point clouds [150]. However, this just opens a number of research paths in this domain, e.g., designing graph convolutional transforms for modeling local statistics

over a point cloud manifold. An important issue in the perception of PC impairments involves gemeotric artefacts. Currently used geometric metrics do not model perceptual phenomena, like the 3D contrast sensitivity and masking. Inspired by mesh fidelity metrics, one could extend these simple geometric metrics to model human visual perception. In addition, it would be an interesting direction to model visual saliency for this kind of content.

Consumer *light-field* cameras, such as the Lytro or Raytrix, are also able to produce a sampling of the plenoptic function, by capturing a high number of micro-images of the scene representing it from multiple points of view. This information, clearly, is highly redundant, and thus requires to be efficiently compressed. Light-field images might be represented in several ways, from macro-pixels to sub-aperture images to epipolar representations. Understanding which representation is more adapt to code and trasmit light-field information is still an open question. A possible approach is to reconstruct the light field based on a sample of the possible views of the scene. This is made possible by the advances on view synthesis using deep learning techniques. Macro-pixel representations, which store light rays arriving at a given spatial point from several directions, might be integrated as attributes into point clouds, potentially solving the main issue of specular reflections — PC texture coding makes the implicit assumption of diffusive surfaces. Finally, a stimulating research direction in immersive communication is *digital holography*, where one of the main limitations is the huge amount of data required to represent high-resolution holograms. It will be interesting to apply similar ideas as those discussed above to this field in the future.

Appendix A

List of publications

Papers in international journals

[J1] A. Rana, G. Valenzise, and F. Dufaux. Learning-based tone mapping operator for efficient image matching. *IEEE Transactions on Multimedia*, 21(1):256–268, January 2019.

[J2] A. Rana, P. Singh, G. Valenzise, F. Dufaux, N. Komodakis, and A. Smolic. Deep tone mapping operator for high dynamic range images. *IEEE Transactions on Image Processing*, 2019.

[J3] Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafal Mantiuk. From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing*, 2019.

[J4] C. Ozcinar, P. Lauga, G. Valenzise, and F. Dufaux. Spatio-temporal constrained tone mapping operator for HDR video compression. *Journal of Visual Communication and Image Representation*, 55:166–178, August 2018.

[J5] N. K. Kottayil, G. Valenzise, F. Dufaux, and I. Cheng. Blind quality estimation by disentangling perceptual and noisy features in high dynamic range images. *IEEE Transactions on Image Processing*, 27(3):1512–1525, March 2018.

[J6] M. Karpushin, G. Valenzise, and F. Dufaux. TRISK: A local features extraction framework for texture-plus-depth content matching. *Image and Vision Computing*, 71:1–16, March 2018.

[J7] W. Fan, G. Valenzise, F. Banterle, and F. Dufaux. Fine-grained detection of inverse tone mapping in HDR images. *Signal Processing*, 152:178 – 188, November 2018.

[J8] E. Zerman, G. Valenzise, and F. Dufaux. An extensive performance evaluation of full-reference HDR image quality metrics. *Quality and User Experience*, 2(1):5, 2017.

[J9] Y. Liu, N. Sidaty, W. Hamidouche, O. Déforges, G. Valenzise, and E. Zerman. An adaptive quantizer for high dynamic range content: Application to video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[J10] V. Hulusic, K. Debattista, G. Valenzise, and F. Dufaux. A model of perceived dynamic range for HDR images. *Signal Processing: Image Communication*, 51:26–39, 2017.

[J11] M. Karpushin, G. Valenzise, and F. Dufaux. Keypoint detection in RGBD images based on an anisotropic scale space. *IEEE Trans. Multimedia*, 18(9):1762–1771, Sep. 2016.

[J12] G. Valenzise, M. Tagliasacchi, and S. Tubaro. Revealing the traces of JPEG compression anti-forensics. *IEEE Trans. Inf. Forensics Security*, 8(2):335 – 349, Feb. 2013.

[J13] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro. No-reference pixel video quality monitoring of channel-induced distortion. *IEEE Trans. Circuits Syst. Video Technol.*, 22(4):605 – 618, Apr. 2012.

[J14] M. Tagliasacchi, G. Valenzise, M. Naccari, and S. Tubaro. A reduced-reference structural similarity approximation for videos corrupted by channel errors. *Springer Multimedia Tools and Applications*, 48:471–492, 2010.

[J15] M. Cossalter, G. Valenzise, M. Tagliasacchi, and S. Tubaro. Joint compressive video coding and analysis. *IEEE Trans. Multimedia*, 12(3):168 –183, Apr. 2010.

[J16] G. Valenzise, G. Prandi, M. Tagliasacchi, and A. Sarti. Identification of sparse audio tampering using distributed source coding and compressive sensing techniques. *EURASIP Journal on Image and Video Processing*, 2009:1–12, 2009.

[J17] M. Tagliasacchi, G. Valenzise, and S. Tubaro. Hash-based identification of sparse image tampering. *IEEE Trans. Image Process.*, 18(11):2491–2504, 2009.

[J18] M. Tagliasacchi, G. Valenzise, and S. Tubaro. Minimum variance optimal rate allocation for multiplexed H.264/AVC bitstreams. *IEEE Trans. Image Process.*, 17(7):1129–1143, 2008.

Papers in proceedings of international conferences

[C1] E. Zerman, G. Valenzise, and A. Smolic. Analysing the impact of cross-content pairs on pairwise comparison scaling. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, June 2019.

[C2] L. Wang, A. Fiandrotti, A. Purica, G. Valenzise, and M. Cagnazzo. Enhancing HEVC spatial prediction by context-based learning. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Brighton, UK, May 2019.

[C3] M. Quach, G. Valenzise, and F. Dufaux. Learning Convolutional Transforms for Lossy Point Cloud Geometry Compression. In *Proc. IEEE Int. Conf. Image Processing*, Taipei, Taiwan, September 2019.

[C4] C. Kang, G. Valenzise, and F. Dufaux. Predicting Subjectivity in Image Aesthetics Assessment. In *21st International Workshop on Multimedia Signal Processing (MMSP'2019)*, Kuala Lumpur, Malaysia, September 2019.

[C5] E. Zerman, V. Hulusic, G. Valenzise, R. Mantiuk, and F. Dufaux. 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)*, Burlingame, USA, January 2018.

[C6] H. Yousef, J. Le Feuvre, G. Valenzise, and V. Hulusic. Video quality evaluation for tile-based spatial adaptation. In *Proc. IEEE Int. Work. Multimedia Signal Processing*, Vancouver, Canada, August 2018.

[C7] G. Valenzise, A. Purica, V. Hulusic, and M. Cagnazzo. Quality Assessment of Deep-Learning-Based Image Compression. In *Proc. IEEE Int. Work. Multimedia Signal Processing*, Vancouver, Canada, August 2018.

[C8] N.K. Kottayil, G. Valenzise, F. Dufaux, and I. Cheng. Learning Local Distortion Visibility from Image Quality. In *Proc. IEEE Int. Conf. Image Processing*, Athens, Greece, October 2018.

[C9] D. Kane, A. Grimaldi, E. Zerman, M. Bertalmío, V. Hulusic, and G. Valenzise. The Preferred System Gamma is Primarily Determined by the Ratio of Dynamic Range of the Original Scene and the Displayed Image. In *IS&T/SPIE Electronic Imaging, Human Vision and Electronic Imaging XXII*, San Francisco, USA, January 2018.

[C10] V. Hulusic, G. Valenzise, and F. Dufaux. Perceived dynamic range of HDR images with no semantic information. In *Human Vision and Electronic Imaging Conference, IS&T International Symposium on Electronic Imaging (EI 2018)*, Burlingame, USA, January 2018.

[C11] E. Zerman, V. Hulusic, G. Valenzise, R. Mantiuk, and F. Dufaux. Effect of color space on high dynamic range video compression performance. In *8th International Workshop on Quality of Multimedia Experience*, Erfurt, Germany, May 2017.

[C12] A. Rana, G. Valenzise, and F. Dufaux. Learning-based tone mapping operator for image matching. In *Proc. IEEE Int. Conf. Image Processing*, Bejing, China, September 2017.

[C13] A. Rana, G. Valenzise, and F. Dufaux. Learning-based adaptive tone mapping for keypoint detection. In *Proc. IEEE Conf. on Multimedia and Expo*, Hong Kong, July 2017.

[C14] Y. Liu, N Sidaty, W. Hamidouche, O. Deforges, G. Valenzise, and Zerman. E. An adaptive perceptual quantization method for HDR video coding. In *Proc. IEEE Int. Conf. Image Processing*, Bejing, China, September 2017.

[C15] M. Karpushin, G. Valenzise, and F. Dufaux. Good features to track for RGBD images. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, New Orleans, USA, March 2017.

[C16] V. Hulusic, G. Valenzise, J. Fournier, J-C. Gicquel, and F. Dufaux. Quality of Experience in UHD-1 Phase 2 television: the contribution of UHD+HFR technology. In *IEEE International Workshop on Multimedia Signal Processing*, London-Luton, United Kingdom, October 2017.

[C17] V. Hulusic, G. Valenzise, K. Debattista, and F. Dufaux. Robust dynamic range computation for High Dynamic Range content. In *Human Vision and Electronic Imaging Conference, IS&T International Symposium on Electronic Imaging (EI 2017)*, Burlingame, USA, January 2017.

[C18] K. Feyiz, F. Kamisli, E. Zerman, G. Valenzise, A. Koz, and F. Dufaux. Statistical Analysis and Directional Coding of Layer-based HDR Image Coding Residue. In *IEEE International Workshop on Multimedia Signal Processing (MMSP 2017)*, London-Luton, United Kingdom, October 2017.

[C19] E. Zerman, G. Valenzise, and F. Dufaux. A dual modulation algorithm for accurate reproduction o high dynamic range video. In *Proc. IEEE Work. on Image, Video and Multidimensional Signal Processing*, Bordeaux, France, July 2016.

[C20] A. Rana, G. Valenzise, and F. Dufaux. Optimizing tone mapping operators for keypoint detection under illumination changes. In *Proc. IEEE Work. on Multimedia Signal Processing*, Montreal, Canada, September 2016.

[C21] A. Rana, G. Valenzise, and F. Dufaux. An evaluation of HDR image matching under extreme illumination changes. In *Proc. IEEE Int. Conf. on Visual Communication and Image Processing*, Chengdu, China, November 2016.

[C22] A. Purica, G. Valenzise, B. Pesquet-Popescu, and F. Dufaux. Using region-of-interest for quality evaluation of DIBR-based view synthesis methods. In *8th Int. Conf. on Quality of Multimedia Experience*, Lisbon, Portugal, June 2016.

[C23] C. Ozcinar, P. Lauga, G. Valenzise, and F. Dufaux. HDR video coding based on a temporally constrained tone mapping operator. In *Proc. IEEE Digital Media Industry and Academic Forum*, Santorini, Greece, July 2016.

[C24] M. Karpushin, G. Valenzise, and F. Dufaux. Keypoint detection in RGBD images based on an efficient viewpoint-covariant multiscale representation. In *Proc. EURASIP European Signal Processing Conference*, Budapest, Hungary, September 2016.

[C25] M. Karpushin, G. Valenzise, and F. Dufaux. An image smoothing operator for fast and accurate scale space approximation. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Shanghai, China, March 2016.

[C26] V. Hulusic, G. Valenzise, E. Provenzi, K. Debattista, and F. Dufaux. Perceived dynamic range of HDR images. In *8th Int. Conf. on Quality of Multimedia Experience*, Lisbon, Portugal, June 2016.

[C27] W. Fan, G. Valenzise, F. Banterle, and F. Dufaux. Forensic detection of inverse tone mapping in HDR images. In *Proc. IEEE Int. Conf. Image Processing*, Phoenix, AZ, USA, September 2016.

[C28] E. Zerman, G. Valenzise, F. De Simone, F. Banterle, and F. Dufaux. Effects of display rendering on HDR image quality assessment. In *Applications of Digital Image Processing XXXVIII, SPIE*, San Diego, CA, USA, 2015.

[C29] A. Rana, G. Valenzise, and F. Dufaux. Evaluation of feature detection in hdr based imaging under changes in illumination conditions. In *Proc. IEEE Int. Symp. Multimedia*, Miami, USA, 2015.

[C30] M. Karpushin, G. Valenzise, and F. Dufaux. A scale space for texture+depth images based on a discrete Laplacian operator. In *Proc. IEEE Int. Conf. Multimedia and Expo*, Turin, Italy, 2015.

[C31] M. Karpushin, G. Valenzise, and F. Dufaux. Improving distinctiveness of BRISK features using depth maps. In *Proc. IEEE Int. Conf. Image Processing*, Quebec City, Canada, 2015.

[C32] G. Valenzise, M. Tagliasacchi, and S. Tubaro. Detectability-quality trade-off in JPEG counter-forensics. In *Proc. IEEE Int. Conf. Image Processing*, Paris, France, 2014.

[C33] G. Valenzise, F. De Simone, P. Lauga, and F. Dufaux. Performance evaluation of objective quality metrics for HDR image compression. In *Applications of Digital Image Processing XXXVII, SPIE*, San Diego, CA, USA, 2014.

[C34] P. Lauga, G. Valenzise, G. Chierchia, and F. Dufaux. Improved tone mapping operator for HDR coding optimizing the distortion/spatial complexity trade-off. In *EUSIPCO*, Lisbon, Portugal, 2014.

[C35] M. Karpushin, G. Valenzise, and F. Dufaux. Local visual features extraction from texture+depth content based on depth image analysis. In *Proc. IEEE Int. Conf. Image Processing*, Paris, France, 2014.

[C36] F. De Simone, G. Valenzise, F. Banterle, P. Lauga, and F. Dufaux. Dynamic range expansion of video sequences: a subjective quality assessment study. In *GlobalSIP*, Atlanta, GA, USA, 2014.

[C37] P. Lauga, A. Koz, G. Valenzise, and F. Dufaux. Segmentation-based optimized tone mapping for high dynamic range image and video coding. In *Proc. IEEE Picture Coding Symposium*, San Jose, California, USA, 2013.

[C38] P. Lauga, A. Koz, G. Valenzise, and F. Dufaux. Region-based tone mapping for efficient high dynamic range video coding. In *Proc. 4th European Workshop on Visual Information Processing*, Paris, France, 2013.

[C39] G. Valenzise, G. Cheung, R. Galvão, M. Cagnazzo, B. Pesquet-Popescu, and A. Ortega. Motion prediction of depth video for depth-image-based rendering using don't care regions. In *Proc. Picture Coding Symposium*, Kraków, Poland, 2012.

[C40] G. Valenzise, M. Tagliasacchi, and S. Tubaro. The cost of JPEG compression anti-forensics. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, 2011.

[C41] G. Valenzise, V. Nobile, M. Tagliasacchi, and S. Tubaro. Countering JPEG anti-forensics. In *Proc. IEEE Int. Conf. Image Processing*, Bruxelles, Belgium, 2011.

[C42] L. Amati, G. Valenzise, A. Ortega, and S. Tubaro. Dependent video coding using a tree representation of pixel dependencies. In *Proceedings of SPIE*, volume 8135, San Diego, California, USA, 2011.

[C43] G. Valenzise, M. Tagliasacchi, and S. Tubaro. Estimating QP and motion vectors in H.264/AVC video from decoded pixels. In *Proc. ACM Int. Workshop on Multimedia in Forensics, Security and Intelligence*, Firenze, Italy, 2010.

[C44] G. Valenzise and A. Ortega. Improved video coding efficiency exploiting tree-based pixelwise coding dependencies. In *Proc. SPIE, Visual Information Processing and Communication*, volume 7543, San Jose, California, USA, 2010.

[C45] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro. Estimating channel-induced distortion in H.264/AVC video without bitstream information. In *Proc. 2nd Int. Workshop on Quality of Multimedia Experience*, Trondheim, Norway, 2010.

[C46] G. Valenzise, M. Tagliasacchi, S. Tubaro, G. Cancelli, and M. Barni. A compressive-sensing based watermarking scheme for sparse image tampering identification. In *Proc. IEEE Int. Conf. Image Processing*, pages 1265 –1268, Cairo, Egypt, 2009.

[C47] P. Tarrío, G. Valenzise, G. Shen, and A. Ortega. Distributed Network Configuration for Wavelet-Based Compression in Sensor Networks. In *Int. Conf. on GeoSensor Networks*, pages 1–10, Oxford, UK, 2009.

[C48] M. Cossalter, M. Tagliasacchi, and G. Valenzise. Privacy-enabled object tracking in video sequences using compressive sensing. In *Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, pages 436–441, 2009.

[C49] A. Albonico, G. Valenzise, M. Naccari, M. Tagliasacchi, and S. Tubaro. A reduced-reference video structural similarity metric based on no-reference estimation of channel-induced distortion. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pages 1857 –1860, 2009.

[C50] G. Valenzise, M. Tagliasacchi, and S. Tubaro. Minimum variance multiplexing of multimedia objects. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA, 2008.

[C51] G. Valenzise, G. Prandi, M. Tagliasacchi, and A. Sarti. Resource constrained efficient acoustic source localization and tracking using a distributed network of microphones.

In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA, 2008.

[C52] G. Valenzise, M. Naccari, M. Tagliasacchi, and S. Tubaro. Reduced-reference estimation of channel-induced video distortion using distributed source coding. In *Proc. ACM Int. Conf. on Multimedia*, Vancouver, Canada, 2008.

[C53] M. Tagliasacchi, G. Valenzise, and S. Tubaro. Localization of sparse image tampering via random projections. In *Proc. IEEE Int. Conf. Image Processing*, San Diego, CA, USA, 2008.

[C54] G. Prandi, G. Valenzise, M. Tagliasacchi, and A. Sarti. Detection and identification of sparse audio tampering using distributed source coding and compressive sensing techniques. In *Proc. 11th Int. Conf. on Digital Audio Effects*, Espoo, Finland, 2008.

[C55] G. Prandi, G. Valenzise, M. Tagliasacchi, F. Antonacci, A. Sarti, and S. Tubaro. Acoustic source localization by fusing distributed microphone arrays measurements. In *Proc. EURASIP European Signal Processing Conference*, Lausanne, Switzerland, 2008.

[C56] A. Bozzon, G. Prandi, G. Valenzise, and M. Tagliasacchi. A music recommendation system based on semantic audio segments similarity. In *Proc. EuroIMSA2008 Internet and Multimedia Systems and Applications*, Innsbruck, Austria, 2008.

[C57] G. Valenzise, M. Tagliasacchi, S. Tubaro, and L. Piccarreta. A ρ-domain rate controller for multiplexed video sequences. In *Proc. Picture Coding Symposium*, Lisboa, Portugal, 2007.

[C58] G. Valenzise, M. Tagliasacchi, and S. Tubaro. A smoothed, minimum distortion-variance rate control algorithm for multiplexed transcoded video sequences. In *Proc. ACM Int. Workshop on Mobile Video*, Augsburg, Germany, 2007.

[C59] G. Valenzise, L. Gerosa, M. Tagliasacchi, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance systems. In *Proc. IEEE Int. Conf. Advanced Video and Signal based Surveillance*, London, UK, 2007.

[C60] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection in noisy environments. In *Proc. EURASIP European Signal Processing Conference*, Poznan, Poland, 2007.

[C61] F. Antonacci, G. Valenzise, L. Gerosa, A. Sarti, and S. Tubaro. Sound-based classification of objects using a robust fingerprinting approach. In *Proc. EURASIP European Signal Processing Conference*, Poznan, Poland, 2007.

Book chapters

[B1] M. Narwaria, M. Perreira Da Silva, P. Le Callet, G. Valenzise, F. De Simone, and F. Dufaux. Quality of experience and HDR: concepts and how to measure it. In F. Dufaux, P. Le Callet, R. Mantiuk, and M. Mrak, editors, *High Dynamic Range Video - From Acquisition, to Display and Applications*. Wiley, 2016.

[B2] M. Narwaria, P. Le Callet, G. Valenzise, F. De Simone, F. Dufaux, and R. Mantiuk. HDR image and video quality prediction. In F. Dufaux, P. Le Callet, R. Mantiuk, and M. Mrak, editors, *High Dynamic Range Video - From Acquisition, to Display and Applications*. Wiley, 2016.

[B3] G. Valenzise, S. Tubaro, and M. Tagliasacchi. Anti-forensics of multimedia data and countermeasures. In T.S. Ho and S. Li, editors, *Handbook of Digital Forensics of Multimedia Data and Devices*. Wiley, 2015.

[B4] E.G. Mora, G. Valenzise, J. Jung, M. Cagnazzo, B. Pesquet-Popescu, and F. Dufaux. Depth video coding technologies. In F. Dufaux, B. Pesquet-Popescu, and M. Cagnazzo, editors, *Emerging Technologies for 3D Video*. Wiley, 2012.

Bibliography

[1] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations," in *Advances in Neural Information Processing Systems*, 2017, pp. 1141–1151.

[2] A. O. Akyüz, R. Fleming, B. E. Riecke, E. Reinhard, and H. H. Bülthoff, "Do HDR displays support LDR content?: A psychophysical evaluation," *ACM Trans. Graph.*, vol. 26, no. 3, 2007.

[3] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, Rhode Island, USA, June 2012.

[4] M. Ashikhmin, "A tone mapping algorithm for high contrast images," in *Proceedings of the 13th Eurographics workshop on Rendering*. Eurographics Association, 2002, pp. 145–156.

[5] T. O. Aydin, R. Mantiuk, and H.-P. Seidel, "Extending quality metrics to full dynamic range images," in *Human Vision and Electronic Imaging XIII*, ser. Proceedings of SPIE, San Jose, USA, Jan 2008.

[6] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *Int. Conf. on Learning Representations (ICLR)*, Vancouver, CA, May 2018.

[7] F. Banterle, K. Debattista, A. Artusi, S. N. Pattanaik, K. Myszkowski, P. Ledda, and A. Chalmers, "High dynamic range imaging and low dynamic range expansion for generating HDR content," *Comput. Graph. Forum*, vol. 28, no. 8, pp. 2343–2367, 2009.

[8] F. Banterle, P. Ledda, K. Debattista, and A. Chalmers, "Inverse tone mapping," in *Proc. Int. Conf. Computer Graphics and Interactive Techniques in Australasia and Southeast Asia*, 2006, pp. 349–356.

[9] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," in *Robotics-DL tentative*. International Society for Optics and Photonics, 1992, pp. 586–606.

[10] T. Bianchi and A. Piva, "Detection of nonaligned double JPEG compression based on integer periodicity maps," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 842–848, 2012.

[11] G. Bjøtegaard, "Calculation of average PSNR differences between RD-curves (vceg-

m33)," VCEG Meeting (ITU-T SG16 Q.6), Austin, Texas, USA,, Tech. Rep. M16090, Apr 2001.

[12] J.-L. Blin, "SAMVIQ - Subjective assessment methodology for video quality," *Rapport technique BPN*, vol. 56, p. 24, 2003.

[13] T. Borer, "Non-linear opto-electrical transfer functions for high dynamic range television," British Broadcasting Corporation (BBC), Tech. Rep. ITU-R WP6C Contribution 369, 2014.

[14] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons." *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[15] M. H. Brill, J. Lubin, P. Costa, S. Wolf, and J. Pearson, "Accuracy and cross-calibration of video quality metrics: new methods from atis/t1a1," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 101–107, 2004.

[16] P. Bronislav, A. Chalmers, P. Zemcík, L. Hooberman, and M. Cadík, "Evaluation of feature point detection in high dynamic range imagery," *Journal of Visual Communication and Image Representation*, pp. 141 – 160, 2016.

[17] M. Brown and D. G. Lowe, "Invariant features from interest point groups." in *Proceed. of British Machine Vision Conf.*, Cardiff, UK, September 2002.

[18] J.-F. Cai, B. Dong, S. Osher, and Z. Shen, "Image restoration: total variation, wavelet frames, and beyond," *Journal of the American Mathematical Society*, vol. 25, no. 4, pp. 1033–1089, 2012.

[19] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 778–792.

[20] G. Cao, Y. Zhao, R. Ni, and X. Li, "Contrast enhancement-based forensics in digital images," *IEEE transactions on information forensics and security*, vol. 9, no. 3, pp. 515–525, 2014.

[21] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, may 2011. [Online]. Available: http://dx.doi.org/10.1007/s10851-010-0251-1

[22] K. Chiu, M. Herf, P. Shirley, S. Swamy, C. Wang, and K. Zimmerman, "Spatially nonuniform scaling functions for high contrast images," in *Proceedings of Graphics Interface '93*, ser. GI '93, Toronto, Ontario, Canada, 1993, pp. 245–253.

[23] P. L. Combettes and J.-C. Pesquet, "Primal-dual splitting algorithm for solving inclusions with mixtures of composite, lipschitzian, and parallel-sum type monotone operators," *Set-Valued and Variational Analysis*, vol. 20, no. 2, pp. 307–330, 2011. [Online]. Available: http://dx.doi.org/10.1007/s11228-011-0191-y

[24] S. Daly, T. Kunkel, X. Sun, S. Farrell, and P. Crum, "41.1: Distinguished paper: Viewer preferences for shadow, diffuse, specular, and emissive luminance limits of high dynamic range displays," in *SID Symposium Digest of Technical Papers*, vol. 44, no. 1. Wiley Online Library, 2013, pp. 563–566.

[25] S. J. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*. International Society for Optics and Photonics, 1992, pp. 2–15.

[26] F. De Simone, G. Valenzise, F. Banterle, P. Lauga, and F. Dufaux, "Dynamic range expansion of video sequences: a subjective quality assessment study," in *GlobalSIP*, Atlanta, GA, USA, 2014.

[27] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proc. SIGGRAPH*, 1997, pp. 369–378.

[28] E. R. do Nascimento, G. L. Oliveira, A. W. Vieira, and M. F. Campos, "On the development of a robust, fast and lightweight keypoint descriptor," *Neurocomputing*, vol. 120, pp. 141–155, 2013.

[29] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive logarithmic mapping for displaying high contrast scenes," *Computer Graphics Forum*, pp. 419–426, 2003.

[30] ——, "Adaptive logarithmic mapping for displaying high contrast scenes," in *Computer Graphics Forum*, vol. 22, no. 3. Wiley Online Library, 2003, pp. 419–426.

[31] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," in *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '02, 2002, pp. 257–266.

[32] ——, "Fast bilateral filtering for the display of high-dynamic-range images," in *ACM transactions on graphics (TOG)*, vol. 21, no. 3. ACM, 2002, pp. 257–266.

[33] M. D. Fairchild, "The HDR photographic survey," in *Color and Imaging Conference*, vol. 2007, no. 1. Society for Imaging Science and Technology, 2007, pp. 233–238.

[34] W. Fan, G. Valenzise, F. Banterle, and F. Dufaux, "Forensic detection of inverse tone mapping in HDR images," in *Proceedings of the International Conference on Image Processing*, Phoenix, AZ, USA, September 2016.

[35] ——, "Fine-grained detection of inverse tone mapping in HDR images," *Signal Processing*, vol. 152, pp. 178 – 188, November 2018.

[36] Z. Fan and R. L. De Queiroz, "Identification of bitmap compression history: Jpeg detection and quantizer estimation," *IEEE Transactions on Image Processing*, vol. 12, no. 2, pp. 230–235, 2003.

[37] E. Francois, C. Fogg, Y. He, X. Li, A. Luthra, and A. Segall, "High Dynamic Range and Wide Color Gamut Video Coding in HEVC: Status and Potential Future Enhancements," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 63–75, Jan 2016. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2015.2461911

[38] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, and H. Brendel, "Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays," in *Proc. SPIE*, 2014.

[39] T. Fujii, K. Mori, K. Takeda, K. Mase, M. Tanimoto, and Y. Suenaga, "Multipoint mea-

suring system for video and sound—100 camera and microphone system," in *IEEE International Conference on Multimedia and Expo*, Toronto, Canada, July 2006.

[40] D. C. Garcia and R. L. d. Queiroz, "Intra-Frame Context-Based Octree Coding for Point-Cloud Geometry," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, October 2018, pp. 1807–1811.

[41] A. Gilchrist, C. Kossyfidis, F. Bonato, T. Agostini, J. Cataliotti, X. Li, B. Spehar, V. Annan, and E. Economou, "An anchoring theory of lightness perception." *Psychological review*, vol. 106, no. 4, p. 795, 1999.

[42] I. Goodfellow, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[43] GPU4vision, "http://gpu4vision.icg.tugraz.at/," Dec 2015. [Online]. Available: http://gpu4vision.icg.tugraz.at/index.php?content=downloads.php

[44] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," *arXiv preprint arXiv:1502.04623*, 2015.

[45] C. Harris and M. Stephens, "A combined corner and edge detector," in *In Proc. of Fourth Alvey Vision Conference*, 1988, pp. 147–151.

[46] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Human vision and electronic imaging VIII*, vol. 5007. International Society for Optics and Photonics, 2003, pp. 87–95.

[47] D. A. Huffman, "A Method for the Construction of Minimum-Redundancy Codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, September 1952.

[48] V. Hulusic, K. Debattista, G. Valenzise, and F. Dufaux, "A model of perceived dynamic range for HDR images," *Signal Processing: Image Communication*, vol. 51, pp. 26–39, 2017.

[49] V. Hulusic, G. Valenzise, K. Debattista, and F. Dufaux, "Robust dynamic range computation for High Dynamic Range content," in *Human Vision and Electronic Imaging Conference, IS&T International Symposium on Electronic Imaging (EI 2017)*, Burlingame, USA, January 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01654136

[50] V. Hulusic, G. Valenzise, and F. Dufaux, "Perceived dynamic range of HDR images with no semantic information," in *Human Vision and Electronic Imaging Conference, IS&T International Symposium on Electronic Imaging (EI 2018)*, Burlingame, USA, January 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01654136

[51] V. Hulusic, G. Valenzise, J. Fournier, J.-C. Gicquel, and F. Dufaux, "Quality of Experience in UHD-1 Phase 2 television: the contribution of UHD+HFR technology," in *IEEE International Workshop on Multimedia Signal Processing*, London-Luton, United Kingdom, October 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01567377

[52] V. Hulusic, G. Valenzise, E. Provenzi, K. Debattista, and F. Dufaux, "Perceived dynamic range of HDR images," in *8th Int. Conf. on Quality of Multimedia Experience*, Lisbon, Portugal, June 2016.

[53] Y. Huo, F. Yang, L. Dong, and V. Brost, "Physiological inverse tone mapping based on retina response," *Vis. Comput.*, vol. 30, no. 5, pp. 507–517, 2014.

[54] ISO/IEC JTC 1/SC 29/ WG 11, "CDVS: Requirements," ISO/IEC, Geneva, MPEG document N11531, July 2010.

[55] ——, "CDVA: Requirements," ISO/IEC, Valencia, MPEG document N14509, March 2014.

[56] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," ITU-R Recommendation BT. 500-13, Jan 2012.

[57] ——, "Methodology for the subjective assessment of the quality of television pictures," *ITU-R Recommendation BT.500-13*, 2012.

[58] ——, "Proposed preliminary draft new Report - Image dynamic range in television systems," ITU-R WP6C Contribution 146, April 2013.

[59] ITU-T, "Parameter values for the HDTV standards for the studio and for international programme exchange," ITU-T Recommendation BT.709, nov 1993.

[60] ——, "Method for specifying accuracy and cross-calibration of video quality metrics (VQM)," ITU-T Recommendation J.149, Mar 2004.

[61] ——, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," ITU-T Recommendation P.1401, Jul 2012.

[62] ——, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," *ITU-T Recommendation P.1401*, 2012.

[63] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Adv. Neural Inf. Process. Syst.*, 1999, pp. 487–493.

[64] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2210–2224, Dec 2015.

[65] B. Jin, M. V. O. Segovia, and S. Süsstrunk, "Image aesthetic predictors based on weighted CNNs," in *IEEE International Conference on Image Processing*. Phoenix, AZ, USA: Ieee, October 2016, pp. 2291–2295.

[66] X. Jin, L. Wu, X. Li, S. Chen, S. Peng, J. Chi, S. Ge, C. Song, and G. Zhao, "Predicting aesthetic score distribution through cumulative jensen-shannon divergence," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[67] C. Kang, G. Valenzise, and F. Dufaux, "Predicting Subjectivity in Image Aesthetics Assessment," in *21st International Workshop on Multimedia Signal Processing*

(MMSP'2019), Kuala Lumpur, Malaysia, September 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02191142

[68] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[69] ——, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2791–2795.

[70] M. Karpushin, G. Valenzise, and F. Dufaux, "Local visual features extraction from texture+depth content based on depth image analysis," in *Proceedings of the International Conference on Image Processing*, Paris, France, 2014.

[71] ——, "Improving distinctiveness of BRISK features using depth maps," in *Proceedings of the International Conference on Image Processing*, Quebec City, Canada, 2015.

[72] ——, "A scale space for texture+depth images based on a discrete Laplacian operator," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Turin, Italy, 2015.

[73] ——, "An image smoothing operator for fast and accurate scale space approximation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, March 2016.

[74] ——, "Keypoint detection in RGBD images based on an anisotropic scale space," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1762–1771, Sep. 2016.

[75] ——, "Keypoint detection in RGBD images based on an efficient viewpoint-covariant multiscale representation," in *Proc. EURASIP European Signal Processing Conference*, Budapest, Hungary, September 2016.

[76] ——, "Good features to track for RGBD images," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, New Orleans, USA, March 2017.

[77] ——, "TRISK: A local features extraction framework for texture-plus-depth content matching," *Image and Vision Computing*, vol. 71, pp. 1–16, March 2018.

[78] M. Karpushin, "Local features for RGBD image matching under viewpoint changes," Ph.D. dissertation, Telecom ParisTech, Paris, France, 2016.

[79] M. Karpushin, G. Valenzise, and F. Dufaux, "A scale space for texture+depth images based on a discrete Laplacian operator," in *IEEE Intern. Conf. on Multimedia and Expo*, Torino, Italy, July 2015.

[80] S. M. Kay, "Fundamentals of statistical signal processing: Detection theory, vol. 2," 1998.

[81] I.-K. Kim, K. McCann, K. Sugimoto, B. Bross, and W.-J. Han, "High efficiency video coding (HEVC) test model 10 (HM10) encoder description," ISO/IEC

JTC1/SC29/WG11, Geneva, Switzerland, Tech. Rep. N12242, Jan. 2013.

[82] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, December 2014, arXiv: 1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[83] P. Korsunov, P. Hanhart, T. Richter, A. Artusi, R. Mantiuk, and T. Ebrahimi, "Subjective quality assessment database of HDR images compressed with JPEG XT," in *7th International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015.

[84] N. K. Kottayil, G. Valenzise, F. Dufaux, and I. Cheng, "Blind quality estimation by disentangling perceptual and noisy features in high dynamic range images," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1512–1525, March 2018.

[85] N. Kottayil, G. Valenzise, F. Dufaux, and I. Cheng, "Learning Local Distortion Visibility from Image Quality," in *Proceedings of the International Conference on Image Processing*, Athens, Greece, October 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01791412

[86] R. P. Kovaleski and M. M. Oliveira, "High-quality reverse tone mapping for a wide range of exposures," in *Proc. SIBGRAPI*, 2014, pp. 49–56.

[87] A. Kovnatsky, M. M. Bronstein, A. M. Bronstein, and R. Kimmel, "Photometric heat kernel signatures," in *Proceed. of 3rd Intern. Conf. on Scale Space and Variational Methods in Computer Vision*, Ein-Gedi, Israel, May 2011.

[88] L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma, "On the accuracy of objective image and video quality models: New methodology for performance evaluation," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2016, pp. 1–6.

[89] M. Krivokuca, M. Koroteev, and P. A. Chou, "A Volumetric Approach to Point Cloud Compression," *arXiv:1810.00484 [eess]*, September 2018, arXiv: 1810.00484. [Online]. Available: http://arxiv.org/abs/1810.00484

[90] J. Kuang, G. M. Johnson, and M. D. Fairchild, "iCAM06: A refined image appearance model for HDR image rendering," *Journal of Visual Communication and Image Representation*, vol. 18, no. 5, pp. 406–414, 2007.

[91] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multi-view imaging and 3DTV," in *IEEE Signal Processing Magazine*, vol. 24, no.6, November 2007.

[92] M. K. Kundu and S. K. Pal, "Thresholding for edge detection using human psychovisual phenomena," *Pattern Recognition Letters*, vol. 4, no. 6, pp. 433–441, 1986.

[93] S. Lasserre, F. LeLéannec, and E. Francois, "Description of HDR sequences proposed by technicolor," *ISO/IEC JTC1/SC29/WG11 JCTVC-P0228, IEEE, San Jose, USA*, 2013.

[94] P. Lauga, A. Koz, G. Valenzise, and F. Dufaux, "Region-based tone mapping for efficient high dynamic range video coding," in *Proc. 4th European Workshop on Visual Information Processing*, Paris, France, 2013.

[95] ——, "Segmentation-based optimized tone mapping for high dynamic range image and video coding," in *Picture Coding Symposium*, San Jose, California, USA, 2013.

[96] P. Lauga, G. Valenzise, G. Chierchia, and F. Dufaux, "Improved tone mapping operator for HDR coding optimizing the distortion/spatial complexity trade-off," in *EUSIPCO*, Lisbon, Portugal, 2014.

[97] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proceed. of IEEE Intern. Conf. on Comp. Vision*, Barcelona, Spain, November 2011.

[98] J. Li, B. Li, J. Xu, R. Xiong, and W. Gao, "Fully connected network-based intra prediction for image coding," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3236–3247, 2018.

[99] J. Li, R. Mantiuk, J. Wang, S. Ling, and P. Le Callet, "Hybrid-MST: A hybrid active sampling strategy for pairwise preference aggregation," in *Adv. in Neur. Inf. Proc. Syst.*, 2018, pp. 3479–3489.

[100] ——, "Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Preference Aggregation," in *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, Montreal, Canada, 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01970953

[101] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. DollÃ¡r, "Focal Loss for Dense Object Detection," *arXiv:1708.02002 [cs]*, August 2017, arXiv: 1708.02002. [Online]. Available: http://arxiv.org/abs/1708.02002

[102] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Processing: Image Communication*, vol. 29, no. 8, pp. 856–863, 2014.

[103] C. Loop, Q. Cai, S. O. Escolano, and P. A. Chou, "Microsoft voxelized upper bodies - a voxelized point cloud dataset," in *ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document m38673/M72012*, May 2016.

[104] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[105] M. R. Luo, G. Cui, and B. Rigg, "The development of the CIE 2000 colour-difference formula: CIEDE2000," *Color Research & Application*, vol. 26, no. 5, pp. 340–350, 2001.

[106] A. Luthra, E. Francois, and W. Husak, "Call for Evidence (CfE) for HDR and WCG Video Coding," ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Tech. Rep. MPEG2011/N12036, Feb 2015.

[107] Z. Mai, H. Mansour, R. Mantiuk, P. Nasiopoulos, R. Ward, and W. Heidrich, "Optimizing a tone curve for backward-compatible high dynamic range image and video compression," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1558–1571, June 2011. [Online]. Available: http://dx.doi.org/10.1109/TIP.2010.2095866

[108] Z. Mai, H. Mansour, P. Nasiopoulos, and R. K. Ward, "Visually Favorable Tone-Mapping With High Compression Performance in Bit-Depth Scalable Video Coding," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1503–1518, Nov 2013. [Online]. Available: http://dx.doi.org/10.1109/TMM.2013.2266633

[109] E. Mair, G. D. Hager, D. Burschka, M. Suppa, and G. Hirzinger, "Adaptive and generic corner detection based on the accelerated segment test," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 183–196.

[110] R. Mantiuk, K. Kim, A. Rempel, and W. Heidrich, "HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions," in *ACM Trans. on Graphics*, vol. 30, no. 4. ACM, 2011, p. 40.

[111] R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "A perceptual framework for contrast processing of high dynamic range images," *ACM Transactions on Applied Perception (TAP)*, vol. 3, no. 3, pp. 286–308, 2006.

[112] R. Mantiuk, K. Myszkowski, and H. P. Seidel, "A perceptual framework for contrast processing of high dynamic range images," *ACM Transactions on Applied Perception*, vol. 3, no. 3, pp. 286–308, July 2006.

[113] R. Mantiuk, S. Daly, and L. Kerofsky, "Display adaptive tone mapping," in *ACM Transactions on Graphics (TOG)*, vol. 27, no. 3. ACM, 2008, p. 68.

[114] W. Mark, L. McMillan, and G. Bishop, "Post-rendering 3D warping," in *Symposium on Interactive 3D Graphics*, New York, NY, April 1997.

[115] R. Mekuria, K. Blom, and P. Cesar, "Design, Implementation, and Evaluation of a Point Cloud Codec for Tele-Immersive Video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 828–842, April 2017.

[116] L. Meylan, S. Daly, and S. Süsstrunk, "Tone mapping for high dynamic range displays," in *Proc. IS&T/SPIE Electronic Imaging: Human Vision and Electronic Imaging XII*, 2007.

[117] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[118] ——, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, 2005.

[119] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[120] S. Miller, M. Nezamabadi, and S. Daly, "Perceptual signal coding for more efficient usage of bit codes," in *Annual Technical Conference Exhibition, SMPTE 2012*, Oct 2012, pp. 1–9. [Online]. Available: http://dx.doi.org/10.5594/M001446

[121] D. Minnen, G. Toderici, M. Covell, T. Chinen, N. Johnston, J. Shor, S. J. Hwang, D. Vincent, and S. Singh, "Spatially adaptive image compression using a tiled deep network," in

2017 IEEE International Conference on Image Processing (ICIP), Sept 2017, pp. 2796–2800.

[122] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[123] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *Signal Processing Letters, IEEE*, vol. 20, no. 3, pp. 209–212, 2013.

[124] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *Signal Processing Letters, IEEE*, vol. 17, no. 5, pp. 513–516, 2010.

[125] ——, "Blind image quality assessment: From natural scene statistics to perceptual quality," *Image Processing, IEEE Transactions on*, vol. 20, no. 12, pp. 3350–3364, 2011.

[126] J.-M. Morel and G. Yu, "ASIFT: A new framework for fully affine invariant image comparison," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 438–469, 2009.

[127] N. Murray and A. Gordo, "A deep architecture for unified aesthetic prediction," *arXiv preprint arXiv:1708.04890*, 2017.

[128] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: a large-scale database for aesthetic visual analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2408–2415.

[129] M. Narwaria, M. P. Da Silva, and P. Le Callet, "HDR-VQM: An objective quality measure for high dynamic range video ," *Signal Processing: Image Communication*, vol. 35, pp. 46 – 60, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596515000703

[130] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pépion, "Impact of tone mapping in high dynamic range image compression," in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Chandler, US, Jan 2014, pp. 1–6.

[131] M. Narwaria, R. K. Mantiuk, M. P. Da Silva, and P. Le Callet, "HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images," *Journal of Electronic Imaging*, vol. 24, no. 1, pp. 010 501–010 501, 2015.

[132] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pepion, "Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality," *Optical Engineering*, vol. 52, no. 10, pp. 102 008–102 008, 2013.

[133] ——, "Tone mapping based HDR compression: Does it affect visual experience?" *Signal Processing: Image Communication*, vol. 29, no. 2, pp. 257–273, 2014.

[134] M. Narwaria, M. P. da Silva, P. Le Callet, G. Valenzise, F. De Simone, and F. Dufaux, "Quality of experience and HDR: concepts and how to measure it," *High Dynamic Range Video: From Acquisition, to Display and Applications*, 2016.

[135] M. Narwaria, C. Mantel, M. Perreira de Silva, and P. Le Callet, "An objective method for High Dynamic Range source content selection," in *6th Int. Work. on Quality of Multime-*

dia Experience, 2014, pp. 13–18.

[136] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[137] C. Ozcinar, P. Lauga, G. Valenzise, and F. Dufaux, "HDR video coding based on a temporally constrained tone mapping operator," in *Proc. IEEE Digital Media Industry and Academic Forum*, Santorini, Greece, July 2016.

[138] ——, "Spatio-temporal constrained tone mapping operator for HDR video compression," *Journal of Visual Communication and Image Representation*, vol. 55, pp. 166–178, August 2018.

[139] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[140] E. Peli, "Contrast in complex images," *JOSA A*, vol. 7, no. 10, pp. 2032–2040, 1990.

[141] M. Perez-Ortiz and R. K. Mantiuk, "A practical guide and software for analysing pairwise comparison experiments," 2017.

[142] M. Perez-Ortiz, A. Mikhailiuk, E. Zerman, V. Hulusic, G. Valenzise, and R. Mantiuk, "From pairwise comparisons and rating to a unified quality scale," *IEEE Transactions on Image Processing*, 2019.

[143] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classifications," in *Proc. European Conf. Comput. Vis.*, 2010, pp. 143–156.

[144] T. Pevný, P. Bas, and J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 2, pp. 215–224, 2010.

[145] T. Pfeiffer, X. A. Gao, Y. Chen, A. Mao, and D. G. Rand, "Adaptive polling for information aggregation." in *The 26th Conference on Artificial Intelligence (AAAI'12)*, 2012.

[146] M. H. Pinson and S. Wolf, "An objective method for combining multiple subjective data sets," in *Visual Communications and Image Processing 2003*. International Society for Optics and Photonics, 2003, pp. 583–592.

[147] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pépion, and P. Le Callet, "Aligning subjective tests using a low cost common set," in *Euro ITV*, 2011.

[148] A. Piva, "An overview on image forensics," *ISRN Signal Processing*, pp. 22 pages,, 2013, Art. ID 496701.

[149] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

[150] M. Quach, G. Valenzise, and F. Dufaux, "Learning Convolutional Transforms for Lossy Point Cloud Geometry Compression," in *Proceedings of the International*

Conference on Image Processing, Taipei, Taiwan, September 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02116891

[151] A. Rana, G. Valenzise, and F. Dufaux, "Evaluation of feature detection in HDR based imaging under changes in illumination conditions," in *Proc. IEEE Int. Symp. Multimedia*, Miami, USA, 2015.

[152] ——, "An evaluation of HDR image matching under extreme illumination changes," in *Proc. IEEE Int. Conf. on Visual Communication and Image Processing*, Chengdu, China, November 2016.

[153] ——, "Optimizing tone mapping operators for keypoint detection under illumination changes," in *Proc. IEEE Work. on Multimedia Signal Processing*, Montreal, Canada, September 2016.

[154] ——, "Learning-based adaptive tone mapping for keypoint detection," in *Proc. IEEE Conf. on Multimedia and Expo*, Hong Kong, July 2017.

[155] ——, "Learning-based tone mapping operator for image matching," in *Proceedings of the International Conference on Image Processing*, Bejing, China, September 2017.

[156] ——, "Learning-based tone mapping operator for efficient image matching," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 256–268, 2019.

[157] E. Reinhard and K. Devlin, "Dynamic range reduction inspired by photoreceptor physiology," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 1, pp. 13–24, Jan 2005.

[158] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," in *ACM Transactions on Graphics*, vol. 21, no. 3. ACM, 2002, pp. 267–276.

[159] ——, "Photographic tone reproduction for digital images," *ACM Transactions on Graphics (TOG)*, vol. 21, no. 3, pp. 267–276, 2002.

[160] A. G. Rempel, M. Trentacoste, H. Seetzen, H. D. Young, W. Heidrich, L. Whitehead, and G. Ward, "LDR2HDR: On-the-fly reverse tone mapping of legacy video and photographs," *ACM Trans. Graph.*, vol. 26, no. 3, 2007, art. ID 39.

[161] T. Richter, "On the standardization of the JPEG XT image compression," in *Picture Coding Symposium (PCS), 2013*. IEEE, 2013, pp. 37–40.

[162] O. Rippel and L. Bourdev, "Real-time adaptive image compression," *arXiv preprint arXiv:1705.05823*, 2017.

[163] A. M. Rohaly, J. Libert, P. Corriveau, A. Webster *et al.*, "Final report from the video quality experts group on the validation of objective models of video quality assessment," *ITU-T Standards Contribution COM*, pp. 9–80, 2000.

[164] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 105–119, 2010.

[165] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *Proceed. of IEEE Intern. Conf. on Comp. Vision*, Barcelona, Spain, November 2011.

[166] R. B. Rusu and S. Cousins, "3D is here: Point cloud library (PCL)," in *Proceed. of IEEE Intern. Conf. on Rob. and Autom.*, Shanghai, China, May 2011.

[167] I. Schiopu, Y. Liu, and A. Munteanu, "CNN-based prediction for lossless coding of photographic images," in *2018 Picture Coding Symposium (PCS)*. IEEE, 2018, pp. 16–20.

[168] K. Schwarz, P. Wieschollek, and H. P. Lensch, "Will people like your image?" *arXiv preprint arXiv:1611.05203*, 2016.

[169] N. Sedaghat, M. Zolfaghari, E. Amiri, and T. Brox, "Orientation-boosted Voxel Nets for 3d Object Recognition," *arXiv:1604.03351 [cs]*, April 2016, arXiv: 1604.03351. [Online]. Available: http://arxiv.org/abs/1604.03351

[170] H. Seetzen, W. Heidrich, W. Stuerzlinger, G. Ward, L. Whitehead, M. Trentacoste, A. Ghosh, and A. Vorozcovs, "High dynamic range display systems," in *Proc. SIGGRAPH*, 2004, pp. 760–768.

[171] G. Sharma and F. Jurie, "Local higher-order statistics (LHS) describing images with statistics of local non-binarized pixel patterns," *Comput. Vis. Image Und.*, vol. 142, pp. 13–22, 2016.

[172] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *Image Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 430–444, 2006.

[173] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *Image Processing, IEEE Transactions on*, vol. 14, no. 12, pp. 2117–2128, 2005.

[174] L. Skorin-Kapov, M. Varela, T. Hoßfeld, and K.-T. Chen, "A survey of emerging concepts and challenges for QoE management of multimedia services," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 2s, pp. 29:1–29:29, 2018.

[175] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, August 2004.

[176] M. C. Stamm and K. R. Liu, "Forensic detection of image manipulation using statistical intrinsic fingerprints," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 492–506, 2010.

[177] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proc. of the Intern. Conf. on Intelligent Robot Systems*, Vilamoura, Algarve, Portugal, October 2012.

[178] C.-C. Su, A. C. Bovik, and L. K. Cormack, "Natural scene statistics of color and range," in *Proceedings of the International Conference on Image Processing*, Brussels, Belgium, 2011.

[179] C.-C. Su, L. K. Cormack, and A. C. Bovik, "Color and depth priors in natural images," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2259–2274, 2013.

[180] V. Sze, M. Budagavi, and G. J. Sullivan, "High efficiency video coding (HEVC)," *Integrated Circuit and Systems, Algorithms and Architectures. Springer*, vol. 39, p. 40, 2014.

[181] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[182] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," in *Int. Conf. on Learning Representations (ICLR)*, Toulon, France, April 2017.

[183] L. L. Thurstone, "A law of comparative judgement," *Psychological Review*, vol. 34, no. 4, pp. 273–286, 1927.

[184] D. Tian, H. Ochimizu, C. Feng, R. Cohen, and A. Vetro, "Geometric distortion metrics for point cloud compression," in *2017 IEEE International Conference on Image Processing (ICIP)*. Beijing: IEEE, September 2017, pp. 3460–3464. [Online]. Available: http://ieeexplore.ieee.org/document/8296925/

[185] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth IEEE International Conference on Computer Vision (ICCV)*, 1998.

[186] I. Tosic and K. Berkner, "3D keypoint detection by light field scale-depth space analysis," in *Proceedings of the International Conference on Image Processing*, Paris, France, October 2014.

[187] A. M. Tourapis and D. Singer, "HDRTools: Software updates," in *ISO/IEC JTC1/SC29/WG11 MPEG2015/M35471, IEEE, Ed., Geneva, Switzerland*, 2015.

[188] G. Valenzise, G. Cheung, R. Galvão, M. Cagnazzo, B. Pesquet-Popescu, and A. Ortega, "Motion prediction of depth video for depth-image-based rendering using don't care regions," in *Proc. Picture Coding Symposium*, Kraków, Poland, 2012.

[189] G. Valenzise, F. De Simone, P. Lauga, and F. Dufaux, "Performance evaluation of objective quality metrics for HDR image compression," in *Applications of Digital Image Processing XXXVII, SPIE*, San Diego, CA, USA, 2014.

[190] G. Valenzise and A. Ortega, "Improved video coding efficiency exploiting tree-based pixelwise coding dependencies," in *Proc. SPIE, Visual Information Processing and Communication*, vol. 7543, San Jose, California, USA, 2010.

[191] G. Valenzise, A. Purica, V. Hulusic, and M. Cagnazzo, "Quality Assessment of Deep-Learning-Based Image Compression," in *Proc. IEEE Int. Work. Multimedia Signal Processing*, Vancouver, Canada, August 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01819588

[192] ——, "Quality assessment of deep-learning-based image compression," in *Multimedia Signal Processing*, Vancouver, Canada, August 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01819588

[193] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proceed. of Intern. Conf. on Multimedia*, ser. MM '10, New York, USA, 2010.

[194] L. Wang, A. Fiandrotti, A. Purica, G. Valenzise, and M. Cagnazzo, "Enhancing HEVC spatial prediction by context-based learning," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Brighton, UK, May 2019.

[195] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[196] Z. Wang, A. Bovik, R. Sheikh, Hamid, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr 2004. [Online]. Available: http://dx.doi.org/10.1109/TIP.2003.819861

[197] Z. Wang and A. C. Bovik, "A universal image quality index," *Signal Processing Letters, IEEE*, vol. 9, no. 3, pp. 81–84, 2002.

[198] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. IEEE, 2003, pp. 1398–1402.

[199] A. B. Watson and L. Kreslake, "Measurement of visual impairment scales for digital video," *SPIE Electronic Imaging, Human Vision and Electronic Imaging VI*, vol. 4299, pp. 79–89, 2001.

[200] N. Werghi, S. Berretti, and A. Del Bimbo, "The Mesh-LBP: a framework for extracting local binary patterns from discrete manifolds," *IEEE Trans. Image Process.*, vol. 24, pp. 220–235, 2015.

[201] C. Wu, B. Clipp, X. Li, J.-M. Frahm, and M. Pollefeys, "3D model matching with viewpoint-invariant patches (VIP)," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, USA, June 2008.

[202] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d ShapeNets: A deep representation for volumetric shapes," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1912–1920.

[203] P. Ye and D. Doermann, "Active sampling for subjective image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4249–4256.

[204] A. Zaharescu, E. Boyer, K. Varanasi, and R. Horaud, "Surface feature detection and description with applications to mesh matching," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami, USA, June 2009.

[205] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser, "3DMatch: Learn-

ing local geometric descriptors from RGB-D reconstructions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2017.

[206] E. Zerman, V. Hulusic, G. Valenzise, R. Mantiuk, and F. Dufaux, "Effect of color space on high dynamic range video compression performance," in *8th International Workshop on Quality of Multimedia Experience*, Erfurt, Germany, May 2017.

[207] ——, "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)*, Burlingame, USA, January 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01654133

[208] E. Zerman, G. Valenzise, F. De Simone, F. Banterle, and F. Dufaux, "Effects of display rendering on HDR image quality assessment," in *Applications of Digital Image Processing XXXVIII, SPIE*, San Diego, CA, USA, 2015.

[209] E. Zerman, G. Valenzise, and F. Dufaux, "A dual modulation algorithm for accurate reproduction o high dynamic range video," in *Proc. IEEE Work. on Image, Video and Multidimensional Signal Processing*, Bordeaux, France, July 2016.

[210] ——, "An extensive performance evaluation of full-reference HDR image quality metrics," *Quality and User Experience*, vol. 2, no. 1, p. 5, 2017.

[211] E. Zerman, G. Valenzise, and A. Smolic, "Analysing the impact of cross-content pairs on pairwise comparison scaling," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, June 2019, pp. 1–6.

[212] G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang, "A psychovisual quality metric in free-energy principle," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 41–52, 2011.

[213] X. Zhang and B. A. Wandell, "A spatial extension of CIELAB for digital color-image reproduction," *Journal of the Society for Information Display*, vol. 5, no. 1, pp. 61–63, 1997.

[214] H. Zhou, T. Sattler, and D. W. Jacobs, "Evaluating local features for day-night matching," in *Computer Vision – ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, 2016, pp. 724–736.

[215] T. Zhu and L. Karam, "A no-reference objective image quality metric based on perceptually weighted local noise," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, pp. 1–8, 2014.