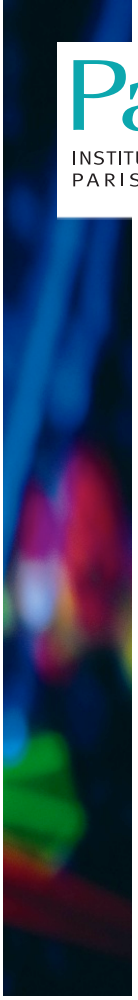
High Dynamic Range (HDR) image analysis

Aakanksha A Rana



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2018-ENST-0015

EDITE - ED 130

Doctorat ParisTech

T H E S I S

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et Images »

présentée et soutenue publiquement par

Aakanksha RANA

le 15 Mars 2018

High Dynamic Range (HDR) Image Analysis

Directeur de thèse : **Frederic DUFAUX**
Co-encadrement de la thèse : **Giuseppe VALENZISE**

Jury

Mme. Sabine SUSSTRUNK, Professeur, Ecole Polytechnique Fédérale de Lausanne	Rapporteur
M. Rafal MANTIUK, Senior Lecturer, University of Cambridge	Rapporteur
Mme. Céline LOSCOS, Professeur, Université de Reims Champagne-Ardenne	Examinatrice
M. Alan CHALMERS, Professeur, University of Warwick	Examinateur
M. Frédéric DUFAUX, Directeur de recherche, CNRS - CentraleSupelec - UPSud	Directeur de thèse
M. Giuseppe VALENZISE, Chargé de recherche, CNRS - CentraleSupelec - UPSud	Co-Encadrant

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

T H È S E

Abstract

High Dynamic Range (HDR) imaging enables to capture a wider dynamic range and color gamut, thus enabling us to draw on subtle, yet discriminating details present both in the extremely dark and bright areas of a scene. Such property is of potential interest for computer vision algorithms where performance degrades substantially when the scenes are captured using traditional low dynamic range (LDR) imagery. While such algorithms have been exhaustively designed using traditional LDR images, little work has been done so far in context of HDR content. In this thesis, we present the quantitative and qualitative analysis of HDR imagery for such task-specific algorithms.

This thesis begins by identifying the most natural and important questions of using HDR content for low-level feature extraction task, which is of fundamental importance for many high-level applications such as stereo vision, localization, matching and retrieval. By conducting a performance evaluation study, we demonstrate how different HDR-based modalities enhance algorithms performance with respect to LDR on a proposed dataset. However, we observe that none of them can optimally to do so across all the scenes. To examine this sub-optimality, we investigate the importance of task-specific objectives for designing optimal modalities through an experimental study. Based on the insights, we attempt to surpass this sub-optimality by designing task-specific HDR tone-mapping operators (TMOs).

In this thesis, we propose three learning based methodologies aimed at optimal mapping of HDR content to enhance the efficiency of local features extraction at each stage namely, detection, description and final matching. By spatial adaptation of a given filter using a regression based approach, we showcase our three models learn to adaptively map the HDR content by bringing invariance to luminance transformations at all the aforementioned stages. We evaluate the performance of all the learning-based models on a proposed HDR dataset of 8 (indoor/outdoor) real scenes where it outperforms existing mapping functions across different feature extraction algorithms.

Finally, this thesis presents three end-to-end deep learning based generic tone mapping (DeepTMOs) designs which cater to desired task-specific characteristics over a wide spectrum of *linear* input HDR images. With the goal of avoiding any specific filtering dependency with a differentiable design as required in previously proposed models, our DeepTMOs serves as a baseline which can be fine-tuned for any computer vision specific task at hand.

Keywords: HDR, Local Features, Computer Vision, Deep Learning, GAN.

Abstract

L'imagerie Haute Gamme Dynamique (HDR) permet de capturer une plage dynamique et une gamme de couleurs plus larges, ce qui nous permet de tirer parti des détails subtils, mais néanmoins distinctifs, prèsents à la fois dans les zones extremement sombres et lumineuses d'une scène. Une telle propriété peut présenter un intérêt potentiel pour les algorithmes de vision par ordinateur lorsque la performance se dégrade considérablement en raison de la perte d'information lorsque les scènes sont capturées à l'aide d'images traditionnelles à faible plage dynamique (LDR). Bien que ces algorithmes aient été conçus de manière exhaustive pour les images LDR traditionnelles, peu de travail a été fait jusqu' à présent dans le contexte du contenu HDR. Dans cette thèse, nous présentons l'analyse quantitative et qualitative de l'imagerie HDR pour de tels algorithmes.

Cette thèse débute par l'identification des questions les plus naturelles et les plus importantes de l'utilisation du contenu HDR pour des tâches d'extraction de caractéristiques de bas niveau, ce qui est d'une importance fondamentale pour de nombreuses applications de haut niveau telles que la vision stéréoscopique, la localisation, l'appariement et la récupération. En réalisant une étude d'évaluation de la performance, nous démontrons comment différentes modalités fondées sur le HDR améliorent la performance des algorithmes par rapport au LDR sur un ensemble de données proposé. Cependant, aucun d'entre eux ne peut le faire de manière optimale dans toutes les scènes. Pour examiner cette sous-optimalité, nous étudions l'importance des objectifs propres à chaque tâche pour concevoir les modalités optimales au moyen d'une étude expérimentale. Sur la base de ces observations, nous tentons de dépasser cette sous-optimalité en concevant des opérateurs de cartographie des tonalités (HDR) spécifiques à chaque tâche.

Dans cette thèse, nous proposons trois méthodologies basées sur l'apprentissage visant à une cartographie optimale du contenu du HDR pour améliorer l'efficacité de l'extraction des caractéristiques locales à chaque étape, à savoir la détection, la description et l'appariement final. Par l'adaptation spatiale d'un filtre donné à l'aide d'une approche par régression, nous présentons nos trois modèles qui apprennent à cartographier de manière adaptative le contenu HDR en apportant une invariance aux transformations de luminance à toutes les étapes susmentionnées. Nous évaluons la performance de tous les modèles basés sur l'apprentissage sur un ensemble de données HDR proposé de 8 scènes réelles (intérieures/extérieures) oú il surpasse les fonctions de cartographie existantes à travers différents algorithmes

d'extraction de caractéristiques.

Enfin, cette thèse présente trois modèles génériques de cartographie tonale (DeepTMO) basés sur l'apprentissage approfondi de bout en bout qui répondent aux caractéristiques spécifiques à la tâche recherchée sur un large spectre d'images HDR d'entrée linéaires. Dans le but d'éviter toute dépendance de filtrage spécifique avec une conception différenciable, comme le requièrent les TMOs proposés précédemment, nos DeepTMOs servent de modèle de base qui peut être affiné pour n'importe quelle tâche spécifique de vision par ordinateur.

Mots clés: HDR, Caractéristiques locales, Vision par ordinateur, Apprentissage approfondi, GAN.

Table of Contents

List of Figures

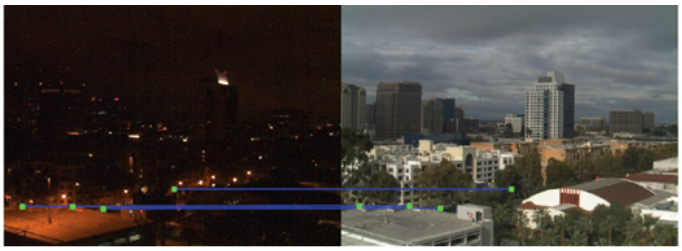
List of Tables

Chapter 1

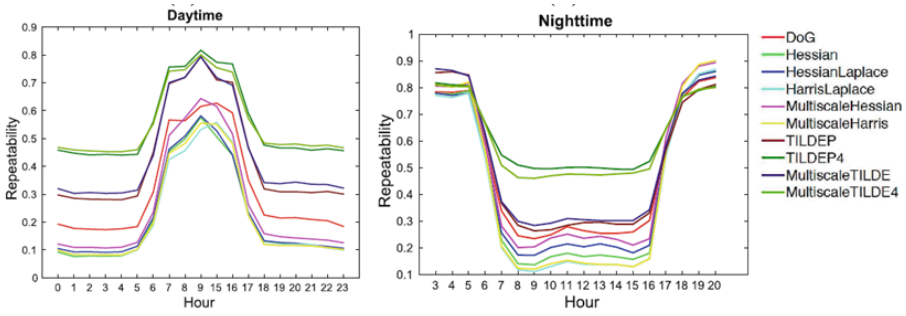
Introduction

High Dynamic Range (HDR) technology has gained immense popularity for its ability to represent a wide range of colors and luminous intensities present in real-world environments [28, 68]. In a sense, these images enable us to draw on subtle, yet discriminating details present both in the extremely dark and bright areas of a scene, which would otherwise get lost in traditional low dynamic range (LDR) imagery. With recent advancements in artificial intelligence, be it in the form of self-driving cars or automated surveillance devices, such high-contrast preserving HDR property is quintessential for the proficiency of underlying computer vision algorithms. In other words, these algorithms should be able to analyze effectively, each and every region in a scene without much uncertainty. Though such algorithms are exhaustively customized for LDR images captured under different conditions, they fail miserably in high-contrast scenes having high or low luminescence [44, 110, 117, 119]. Since high-contrast scenes are extremely common in the real world, it becomes quite critical in cases like automated vehicles where human lives are involved. Thus, it necessitates the application of HDR technology for viability of computer vision algorithms. While several algorithms have been exhaustively designed for interpreting over or under exposed scenes using LDR images, little work has been done so far in context of HDR content.

This thesis is focused on the analysis of 'enriched' HDR images for the benefit of low-level visual features correspondence problem, which is the bedrock for many other high-level computer vision algorithms including, registration and stereo vision, motion estimation and localization, matching, retrieval and recognition of objects and actions. More specifically, the thesis investigates the fundamental challenges involved in using HDR imaging and derives the optimal ways of using HDR content for enhancing the robustness of such tasks.



(a) Day-Night Matching Example



(b) Repeatability Rate of detection of Keypoints

Figure 1.1 – (a) shows an example case from [119], where matching of salient points (common in both) is shown using blue lines between two images of same scene taken at different hours of the day. (b) shows the efficiency measure Repeatability rate RR over a large dataset of LDR day/night images using the state-of-the-art techniques. The crest and the troughs in the curves illustrates that image captured during the daylight matches well with only other day-time images and not with the ones captured in the dark.

1.1 Context and Objectives

The robustness of computer vision applications can be construed from a three-level feature hierarchy namely, low-level, mid-level and high-level. Since the latter two levels build heavily from the former, the efficacy in low-level analysis has been considered quintessential [116]. Generally, the low-level analysis is defined and evaluated in terms of 'visual features correspondence' problem [96]. The problem is formulated by drawing correspondence between images using *visual features extraction* algorithms. Visual features are the discriminative signatures that contain local information from the salient locations in the images. The correspondence between these features defines the 'match-ability' between the two contents. An example from [119] is shown in Figure 1.1 (a), depicting the correspondence between a day and night scene using LDR images.

Several attempts including local [44, 103, 112], global [97] normalization models and learning-based method [110, 116], have been made to ensure better luminance invariant designs in LDR imagery. However, these techniques are practically inefficient to completely compensate the loss of information or comprehend the change in spatial configurations of objects present in a scene. Consequently, these algorithms fail to find true correspondences between similar objects and result in sharp decline in performance. An example [119] of day/night matching is shown in Figure 1.1, where in the performance scores of state-of-the-

art feature detection algorithms drops significantly in day/night lighting variations.

HDR imaging on the other hand, can partially overcome such limitations by capturing a wide range of radiance and luminosity while preserving fine details in both 'dark and overly bright' regions. Hence, owing to its extended capabilities, use of HDR imagery in local feature extraction is essential.

Local feature extraction algorithms have been extensively explored in the computer vision literature. All these algorithms have been designed and optimized with respect to LDR images. These images store gamma-encoded values [0,255] and are generally represented using an 8-bit integer representation. On the contrary, HDR pixels are real valued and proportional to the physical luminance of the scene, expressed in cd/m^2 and can vary up to $10^5 cd/m^2$ on a sunshine day [91]. Consequently, HDR images have largely varying pixel intensities. Hence, it raises a natural question of how to begin with the HDR image analysis for feature extraction algorithms. In simple words, it is not clear whether HDR images can be directly used with such algorithms.

One alternative would be to optimize each feature extraction pipeline for HDR images. But it would be quite impractical and cumbersome specifically for existing learning-based pipelines which would require a large amount of geometrically-calibrated HDR dataset. Not to mention, it might complicate the direct plug-in possibilities with existing mid-level and high-level computer vision pipelines.

In this thesis, we hence opt for other solution. *We concentrate on the HDR images on the input side and explore which is the best way to employ such images in LDR-optimized feature extraction algorithms.*

Some HDR based studies [2, 11] have recently investigated the impact of using HDR images on features detection performance. Since the algorithms are LDR-optimized, they first convert the HDR content into an LDR image using some Tone-Mapping Operators(TMOs) and then apply feature detection techniques. These studies however, do not explore other modes of using HDR (such as linear) and also lacks the impact of using different varieties of existing TMOs.

Research in HDR imagery has always been addressed from a perceptual perspective point of view. Therefore, all modes of using HDR, referred as 'modalities' in this thesis, have been accustomed to human-vision attributes [13] *e.g.* preserving image aesthetics, contrast etc. One common way of assessing the HDR content is via tone mapping. By definition, TMOs are the models aiming to map HDR content in a suitable 8-bit LDR representation for displaying content on standard display screens. For instance, a popular technique involves the compression of estimated luminance e.g., using edge preserving filters such as bilateral [29] from HDR scenes in order to produce a visually pleasing tone-mapped output.

Conceptually, perceptual objectives are quite unrelated to the task-specific performance criteria such as precision score for feature correspondence. Unlike visual perception, the feature extraction pipelines follow strict designs to develop invariance in sparsely located

pixel-level information such as *histogram of gradient* orientations. Therefore, even though tone mapped images are LDR images with better contrast, it is rather questionable if they are *optimal* to extract robust visual features.

Therefore, *it remains unclear which is the best way to employ HDR; linear HDR images, some other form of encoded HDR or quantizing the information somehow into another LDR representation e.g.*using the TMOs ?

Based upon these considerations, this thesis begins to explore the naturally raised questions from scratch, since it has not yet been investigated in literature: **1) what specific advantages can HDR images bring to the existing feature extraction pipelines quantitatively and qualitatively, compared to the existing LDR approaches? 2) what are the best possible ways of using HDR content ?**

Attempting to answer the aforementioned questions, in this thesis we proceed with the following step-wise methodology:

1. We first investigate the HDR and its several corresponding modalities using state-of-the-art feature extraction approaches subject to their covariance and invariance under drastic lighting variations. By proposing a geometrically calibrated dataset, the analysis of various forms of HDR inputs is based on an elaborated study of their performance on two key feature stability stages: (i) keypoint detection, (ii) descriptor extraction.

2. Based on state-of-the-art techniques of local feature extraction, we design the optimal methodologies to use HDR content aiming at facilitating stable and efficient correspondences between grayscale images in real-world luminance conditions, which are quite challenging for traditional LDR images. These luminance conditions comprise of substantial changes in terms of day-night lighting variations in outdoor scenes and changes in position of strong reflectors resulting in highly saturated regions in indoor scenes.

3. We evaluate the performance of the proposed optimal methodologies of using HDR content and compare them to the state-of-the-art approaches in image matching scenarios together with a standard repeatability and feature distinctiveness measures.

1.2 Contributions

The following contributions, mainly to the field of HDR image analysis, are presented in this thesis. The published articles are reported in List of Publications in Section 7.2.

1. We investigate how much gains can HDR bring over LDR for the keypoint detection task, and which are the best modalities of using HDR to obtain it. To this end, we additionally capture a dataset with two scenes having a wide range of illumination conditions. This contribution is has been presented in the following article:

A. Rana and G. Valenzise and F. Dufaux, "Evaluation of Feature Detection in HDR Based Imaging Under Changes in Illumination Conditions", IEEE International Symposium on Multimedia (ISM), *Miami,USA, December, 2015.*

2. We evaluate the performance of various HDR and LDR modalities for full feature extraction pipeline, including detection and description individually. We show that since the two step are independent, HDR representations that work best for keypoint detection are not necessarily optimal when full feature extraction is taken into account. The following paper details this work:

   A. Rana and G. Valenzise and F. Dufaux, "An Evaluation of HDR Image Matching under Extreme Illumination Changes", The International Conference on Visual Communications and Image Processing (VCIP), *Chengdu, China, 2015.*

3. We discuss the sub-optimality of existing TMOs and what is needed to design a keypoint-optimized TMO. To that end, we draw comparison between the optimization of existing TMO parameters with respect to: a) task-specific measure *i.e.* Repeatability Rate RR and b) statistical correlation coefficient CC between pairs of tone-mapped images of the same scene with lighting variations. CC measures the statistical similarity between a pair of tone-mapped images. This is presented in the following paper:

   A. Rana and G. Valenzise and F. Dufaux, "Optimizing Tone Mapping Operators for Keypoint Detection under Illumination Changes", 2016 IEEE Workshop on Multimedia Signal Processing (MMSP 2016), *Montréal, Canada, 2016.*

4. We design a learning-based adaptive tone mapping framework which aims at enhancing keypoint stability by design a pixel-wise adaptive TMO. The regression based model is driven by Support Vector Regression (SVR) using keypoint characteristics. Additionally, we propose a simple detection-similarity-maximization model to generate appropriate training samples. We present this contribution in the following paper:

   A. Rana and G. Valenzise and F. Dufaux, "Learning-based Adaptive Tone Mapping for Keypoint Detection", The International Conference on Visual Communications and Image Processing(ICME), *Hong Kong, China, 2017.*

5. We propose to optimally tone-map a high dynamic range (HDR) content for invariant *descriptor extraction* under drastic illumination variations. We employ a learned model to predict optimal modulation maps that help to locally alter the intrinsic characteristics (such as shape, size) of the tone mapping function. The detail of this work is available in the following paper:

   A. Rana and G. Valenzise and F. Dufaux, "Learning-based Tone Mapping Operator for Image Matching", IEEE International Conference on Image Processing (ICIP'2017), *Beijing, China 2017.*

6. We address the sub-optimality of TMOs by collectively addressing both stages of keypoint detection and descriptor extraction in the feature matching framework. We develop a two-step framework, consisting of: a) a luminance-invariant guidance model based upon a Support Vector Regressor (SVR) to optimally adapt the tone mapping function for image matching; and b) an energy maximization model to generate appropriate training samples considering each independent proxy function. The article describing this work is currently under revision:

   A. Rana and G. Valenzise and F. Dufaux, "Learning-based tone mapping operator for efficient image matching", IEEE Transaction of Multimedia(TMM)*, 2017 accepted*

7. We propose a deep learning based Tone mapping operator which predicts high quality tone mapped outputs over a wide spectrum of *linear* HDR images. The proposed model is designed for perceptual objectives. However, this is the first end-to-end learnable TMO which can be fine-tuned for any computer vision specific task such as image matching. The following article describing this work is submitted:

   A. Rana, P. Singh*, G. Valenzise, F. Dufaux and N. Komodakis. "Deep Tone Mapping Operator for High Dynamic Range Imagery",* ACM SIGGRAPH*, 2018 submitted.*

1.3 Structure of the thesis

This thesis is structured into 7 chapters

- Chapter 2 discusses the background of HDR imagery using conventional acquisition-generation-display approaches and brief history of HDR imagery for computer vision applications. The chapter provides details about benchmark studies on fundamental visual feature extraction algorithms. The quantitative evaluation metrics for accessing the feature extraction and HDR imaging, used in following chapters, is revisited in detail.

- Technical contributions of the thesis begin in Chapter 3, by addressing the performance evaluation of the different HDR formats for two stage of local feature extraction, which is the keypoint detection and description. We propose a geometrically calibrated HDR luminance change dataset. In this chapter, we investigate how different HDR formats can impact the keypoint detection and descriptor performance and if there is any format which yields stable keypoints across these scenes. Specifically, we compare 11 image formats and test the keypoint detection and full feature extraction efficiency on them using two different detectors.

- Based on preceding observations, in Chapter 8 we propose a learning based adaptive tone mapping framework for HDR images which results in stable and efficient keypoint

detection. In this chapter, we initially present an experimental study showing what it takes to optimize a tone mapping function for a metric task such as keypoint detection. Later, we present a regression based guidance model to predict the desired pixel-wise modulation maps by using the linear HDR content from scenes captured with varying lighting conditions.

- After addressing the keypoint detection, in Chapter 5, we move to full feature extraction. To this end, the chapter first proposes a descriptor-optimal TMO design which solely aims at the extraction of invariant (as much as possible) descriptors from high-contrast areas of the scenes. Later, an optimal TMO OpTMO for full feature extraction chain (including both detectors and descriptors) is introduced which simultaneously enhances the detection rates and matching of features by inculcating proxy cost functions; to fuse the relevant information from independent design objectives.

- In Chapter 6, we design the first end-to-end deep learning based TMO. Being trained with a perceptual objective in its primal stage, the GAN based network defines a universally applicable tone mapping function which yields most natural images. The proposed model can be simple fined tuned with any desired objective such as image matching and eradicates the need of designing any proxy cost functions. Instead, it provides a baseline architecture to explore HDR imagery for several other domain specific analysis tasks such as medical image analysis or high resolution remote sensing tasks.

- Finally, in Chapter 7, we present concluding remarks and briefly describes future work perspectives.

Chapter 2

Background and State of the Art

High Dynamic Range (HDR) imaging has been a subject of interest in graphics community over the past decades, inspiring to capture and reproduce a wide range of colors and luminous intensities of real world on a digital canvas. Primarily, the research in this domain started with problems in generation, acquisition and display. An excellent brief overview can be found in [14, 68]. A simple method of HDR capture involves multiple exposure images of the same scene taken at different time-exposure settings. To display such scenes on standard display screen, a variety of Tone Mapping Operators (TMOs) have been designed, promising the most honest representation of real world luminosity and color gamut. Several processing methodologies such as de-ghosting [24] have also been addressed to provide a refined and artifact-free HDR representation. However, HDR imagery has not been fully explored for computer vision problems such as feature extraction.

In this chapter, we first briefly describe the classical way of handling HDR content, where and how it has been used in the computer vision applications. Then, we provide the details of feature extraction techniques which have been addressed throughout the thesis.

2.1 HDR Imaging

Our real-world scenes are much more brighter and colorful, and contains higher contrast than what is reproduced in 8-bit LDR images. Unlike these traditional technologies, HDR imaging is a technology which represents the wide range of colors and luminosities available in real world scenes in the form of digital images and videos. The luminance information in HDR images is generally represented using floating point formats that can use up to 32 bits, differently from traditional 8-bit LDR formats that store gamma-encoded values (approximately linear to perception). As a consequence, HDR images stores vast range of information in the dark as well as bright regions of the scene.

One most commonly adopted method to generate HDR images is by capturing multiple LDR pictures of a scene at different exposure times, in order to estimate a signal proportional to the physical luminance of the scene [24, 73] as shown in Fig 2.1. After computing a

Figure 2.1 – Taken from [24]. Multiple exposures of a Church scene along with final Radience Map.

camera response curve and normalizing by the exposure change, we obtain a single HDR image by weighted averaging of pixel values across these different exposures. While various weighting strategies have been proposed in the literature, we adopted the setting proposed in [24] for capturing our dataset.

2.1.1 HDR imaging for Display

Conventional display technology assume that the input image is LDR. In order to compress the dynamic range of an HDR image to LDR, a great variety of TMOs addressing different perceptual objectives have been proposed in the past years.

Tone mapping operators have been classified into several categories principally based upon how they handle the contrast, color and luminosity in a given HDR scene [7]. Overall, these algorithms have been classified into *global* and local approaches. The global methods such as [27, 58, 95] apply the same compression function to all the pixels of an image. For the *local* techniques such as [21, 79, 105], a tone-mapped pixel depends on the values of neighboring pixels. Even though global approaches are faster to compute, their resulting LDR outputs do not maintain adequate contrast in the images and thus the scene appears somewhat washed out. The local tone mapping functions, conversely do not face these issues and are generally capable of handling contrast ratios, meanwhile preserving the details. However, these operators result in some prominent 'halo' effects around the high frequency edges, thereby giving unnatural artifacts in the scenes. Another kind of function is a perceptual mapping operator [29, 35, 67] which, inspired from the human visual system, models attributes such as adaptation with time, discrimination at high contrast stimuli and gradient sensitivities. Although these methods yield detailed outputs at high computational cost, the aesthetic appeal of generated images is questionable. All these tone mapping methods have aimed to produce images which are quite close to what an individual would perceive in reality. The performance evaluation of these TMOs have been widely studied only from a perceptual point of view [13, 59] and TMOs, generally, has been used only for display applications.

Fine tuning of parameters to enhance the perceived visual quality of tone mapped image has been previously explored in the TMO literature [7, 91]. Mostly, such parameters

were tuned either by a trial-test or grid-search based approach to yield favorable outputs for a wide variety of scenes [21, 29, 67, 90]. Although some works even propose to automate the parameter selection [89], the tuned values are applied globally over the scene.

2.1.2 HDR Imaging for Computer Vision Applications

Literature of HDR imaging applied to computer vision problems is not very vast. It is only recently that HDR imaging has been considered in the computer vision applications such as local feature analysis [12, 55], video surveillance [9, 49] and photogrammetric applications [101]. In the following, we briefly describe various applicative scenarios where HDR imagery has been proposed to enhance the task specific performances.

1. **Local features**: Considering both detection and description stages, in [19, 20], the added value of using HDR video has been studied in the context of matching in outdoor locations, as well as pedestrian and vehicle tracking. In [20], authors compared the feature matching performance of SURF and SIFT descriptors using a dataset of both indoor and outdoor HDR (16-bit) images. Another interesting work has been carried out by Pribyl et al. [11], where authors presented an evaluation of the repeatability of state-of-the-art keypoint detectors on images under different transformations (lighting, viewpoint, distance) for different LDR/HDR modalities, including simple global and local TMO's but *not* the original HDR values. Only based on description assessment, [22] presented a normalization approach, where TMO is used to remove lighting-dependent information from an HDR picture, and leaving only the object's texture. In [22], results are shown in terms of SIFT descriptors matching performance, on a limited dataset of two images, in comparison to two popular TMO's.

2. **Tracking and Video Surveillance**: HDR imaging has attracted interest in the field of video surveillance. Early work on analyzing TMO for surveillance applications was carried out by [9], who propose to combine the properties of local and global TMO's for object detection and tracking. However, that work lacks the comparison in terms of detection accuracy with other TMO's. In [2], an interesting scenario of enhanced people detection and tracking in indoor HDR scenes is presented using only one sequence.

3. **Photogrammetric Applications**: Suma et al. [101] presented the added value of using HDR imagery and evaluated the performance of different TMOs in the context of photogrammetric applications by estimating the enhanced count of features in different TMOs over LDR. [55] made similar investigation with the enhanced number of local invariant features on detailed architectural scenes in HDR over LDR images. Note that the number of detected feature points is not itself a sufficient indicator of detection performance.

4. **Face detection**: In [56], authors evaluated TMOs for face recognition applications based on the subjective and the objective evaluation metrics. [69] investigated the power of high contrast tone mapped images for automatic face recognition using sparse representation techniques.

5. **Privacy Protection**: Rerabek et al. [92] considered the implications of having HDR content on privacy protection; a subject of great practical interest in surveillance scenarios.

One commonality amongst all these studies is the use of existing perception-based TMOs. These techniques have been directly used to convert HDR images to LDR. It is difficult to contemplate whether the adopted procedure is optimal or not. In Chapter 3, we discuss this question with experimental evaluations for feature extraction algorithms. Additionally, most of the studies were tested on small sets of 1 or 2 scenes.

2.2 Local Visual Features

Feature extraction algorithms play a critical role in several computer vision pipelines. As discussed in Chapter 1, most of these algorithms are optimized for LDR imagery. Literature confining these algorithms is immense but essentially revolves around its two stages, namely keypoint detection and descriptor extraction. Keypoint detection methods look for covariant salient locations in a scene that can be repeatedly detected when the scene is undergoing drastic geometrical and photometric transformations [45, 64, 96]. Later, descriptor extraction algorithms are applied to extract discriminative invariant signatures from these selected keypoint locations [64, 70, 116].

Over the past couple of decades, these algorithms have shifted the course from hand-crafted to learning based mechanisms, but the competitiveness of some classical methods such as SIFT, SURF can be observed over the vast range of transformations [119]. In the following, we discuss various detection and descriptor extraction algorithms which we have adopted to design the evaluation framework throughout the thesis.

2.2.1 Keypoint Detection

Algorithms for keypoint detection, in general, have been categorized in corner [45, 110] and blob detectors [70]. We discus the principles in designing the corner and blob detectors, in the following.

The concept of *corner-like* keypoint detection methods has gained popularity for low-latency vision tasks due to high speed, less computational complexity and competitive accuracy [107]. By definition, corners exhibit low correlation with neighboring pixels in all directions. The most basic and widely adopted corner detectors [36, 45, 104] localize the extrema primarily in an image I by computing the per pixel autocorrelation matrix or the

structural matrix given as

$$\boldsymbol{M} = \begin{bmatrix} I_x^2 & I_{xy} \\ I_{yx} & I_y^2 \end{bmatrix}, \tag{2.1}$$

where each component represent the directional derivative. Thereafter, different methods are proposed in the literature to localize the extrema 'keypoint' [96]. [45] describes the response $\mathcal{R}$ point score for each pixel x by computing the associated eigenvector and the directional intensity variation, given as:

$$\mathcal{R}(\boldsymbol{x}) = det\{\boldsymbol{M}(x)\} - k \cdot tr\{\boldsymbol{M}(x)\}^2, \tag{2.2}$$

where k is tuned empirically. This method is not only efficient in practice for real-time corner detection, but also optimal for locating center of junctions and circular symmetric structures [45]. For [104], the algorithm relies on the same aforementioned second moment matrix M, but explicitly computes its eigenvalues different from the previous Harris detector using the following function

$$\mathcal{R} = min(\lambda_1, \lambda_2). \tag{2.3}$$

Although this enhances the computational requirements, the feature points detected are better localized.

Although the corner detectors are computationally fast and robust to variations such as translation and rotation, their designs are weak to handle scale variations. Additionally, since most of them are located on the object boundaries, corner points are prone to failure with scene content changing its spatial configuration such in lighting variations. To address this, blob detection methods have been designed to detect regions. Blob detectors have been discussed in details in [70]. In this thesis, we have mainly focused on widely adopted SIFT and SURF methods to evaluate our models.

SIFT [64] is one of the first and most commonly used blob detectors which is based on the principles of Laplacian of the Gaussian (LoG) (L). The algorithm is considered invariant to scale, rotation, illumination and viewpoint. For a given image I, the convolution take place at different scales t using the Gaussian kernel given as

$$G(x,t) = \frac{1}{2\pi t} \exp^{\frac{-\|x\|^2}{2t}} \tag{2.4}$$

and then, the multi-scale image Gaussian pyramid is obtained. Since LoG is computationally expensive, one simple method to approximate it, is by computing the difference of two close levels at each given scale level of multi-scale pyramid. The proofs are described in [64]. Generally, the approximation is given as

$$\nabla^2 L(x,t) \approx \frac{t}{\delta t}(L(x, t-\delta t) - L(x,t)) \tag{2.5}$$

where x is the pixel location. It often referred to as DoG (Difference of Gaussians) operator

which looks for distinctive blobs or regions. Later, to locate the maxima and minima for keypoint localization a simple min-max suppression technique is applied iteratively through each pixel while checking all neighbor pixels.

Although, DoG has been introduced for reducing the computational complexity, the practicality of SIFT has always been questioned for real-time tasks such as tracking. Therefore, the SURF detector has been introduced in [8] to compensate the computational requirement. It uses an integer approximation of the determinant of Hessian which is computed on different layers of a multi-scale representation. Different from SIFT, the SURF detectors rely on bi-directional Gaussian filtering to reduce the time-complexity.

Another category of detector include FAST [94] and BRISK [61]. The basic strategy for both methods is inspired from corner like detectors, however, the underneath principle is entirely different. In FAST, a pixel is considered a keypoint if its N (value is 16) contiguous surrounding pixels (on a circle) are either brighter or darker then its intensity value. This detection method is considered to be the first learning based method, even though learning techniques have been used simply to speedup the localization process. BRISK on the other hand is a an extension detection scheme satisfying the covariance to scale and rotation by using a multi-scale image representation.

These methods are computationally fast and are widely used for real time applications such as object localization and tracking.

2.2.2 Descriptor Extraction

The descriptor extraction algorithms have gone hand-in-hand with the keypoint detection literature and have been thoroughly studied (see, e.g., [70]). The main goal is to represent the visual information present in a local image patch in the form of a unique invariant signature which can help to find its true correspondence in other images. In this thesis, we consider the following four widely accepted feature extraction schemes as namely: BRISK [61] and FREAK [77] (corner based), SIFT [64] and SURF [8] (blob based) to evaluate the performances of our models.

BRISK [61] is a computationally efficient scheme which is made up of a fast multi-scale detector and a binary descriptor. Its detection module is an extension of corner-based detectors such as FAST or Harris as explained in the Keypoint detection section. The BRISK descriptor is a binary string computed by brightness comparisons on circular sampling patterns around the detected regions. These descriptors are binary in nature and have low storage costs. To compute the distances between the descriptors, Hamming distance metrics are used which is faster than the Euclidean distance metric.

Another binary feature extraction scheme employed for evaluation in this thesis is FREAK [77]. It is composed of a Harris corner detector and a binary descriptor. Similar to BRISK descriptor, FREAK also uses a concentric rings arrangement, but the sampling grid is non-uniform as inner circular rings have exponentially more points. Hamming distance

metric is used to compute the distances between the descriptors.

The third feature extraction scheme is SIFT [64] which is a classical algorithm consisting of a blob keypoint detector (based on difference of Gaussians) and a gradient-based descriptor. Its DoG based detector is detailed in the keypoint detection section. The descriptor part is a 128-dimensional histogram which is formed by concatenation of the image gradients computed on 4×4 grid spatial neighborhood around the detected keypoint. The mathematical notations regarding the SIFT descriptors are detailed in Chapter 5.

Lastly, we discuss about the SURF [8] feature extraction scheme. It is composed of a computationally efficient blob type detector mainly based on the Hessian matrix approximation as discussed in keypoint detection section. Its descriptor is computed as the sum of the Haar wavelet response around the point of interest. To compute the SURF descriptor, firstly the region of interest is structured into 4x4 grids. Then, the Haar wavelet responses are computed from 5x5 sampled points for each grid, with a spatial Gaussian weighting.

2.2.3 Performance Evaluation Metrics

Keypoint detection and descriptor extraction performance on the LDR images are measured using the standard criteria of Repeatability Rate (RR) and Matching Score (MS) respectively, as detailed in [70, 71]. Whereas for the evaluation of the full image matching, the mean average precision (mAP) scores [70] are usually computed. In this thesis, we compare our proposed models with the tone-mapping models by computing their performances using these feature extraction metrics. In the following, we detail RR, MS and mAP metric.

- *Keypoint accuracy measure*: RR is the most common and widely used measure of detector efficiency. In mathematical notations, it is defined as

$$\frac{r_{ref}(\epsilon)}{\min(n_{ref}, n_{test})}, \tag{2.6}$$

  where r_{ref} is the number of keypoints detected in the reference image which are *repeated* in the test image, and n_{ref} and n_{test} are the number of detected keypoints in the reference and test image, respectively. A keypoint is considered to be *repeated* in the test image if: a) it is detected as a keypoint in the test image, and b) it lies in a circle of radius ϵ centered on the projection of the reference keypoint onto the test image. Generally, the number of detected keypoints in images vary at large which might lead to a bias in the final RR. Therefore, in our evaluation framework, we select the strongest N number of keypoints to avoid any form of numerical bias.

- *Descriptor Matching*: The match-ability of a decriptor is measured using the MS. It is defined as the fraction of correct matches to the total number of correspondences in the image pair. A match has been defined using three different matching strategies

in [70] namely threshold based matching, nearest neighbor (NN) matching and nearest neighbor distance ratio (NNDR). In threshold-based matching two descriptors are said to have a match if the distance between them is below a certain threshold. As a descriptor can have several matches, this strategy is error-prone and is seldom used. In nearest neighbor (NN) matching, a descriptor A finds its match B only if A is the nearest neighbor to B and if the distance between them is below a threshold. Nearest neighbor distance ratio (NNDR) extends NN by introducing a threshold to the ratio of the distance descriptors. More precisely, a descriptor finds a good match if the ratio between its distance from the first closest match and its distance from the second closest match is less than a given threshold th. These distances depend on the descriptor type, *i.e.*Hamming distance metric is used for binary descriptors and Euclidean distance is used for non-binary descriptors.

To define a correct match, feature location is taken into account. Two descriptors yield a true positive match if they correspond to two keypoints/regions which are repeated [70] in the reference and query images. Similarly, a match is labeled as a false positive if the corresponding keypoints are not repeated.

In summary, only nearest neighbor (NN) and nearest neighbor distance ratio (NNDR) can reduce the possibility of one-to-many matching scenarios of the descriptors. In most of our evaluation frameworks, we employed the NNDR matching strategy to compare the performance of our TMO with other techniques.

- *Feature Matching Efficiency*: MS gives only the estimate of correct matches, while in practice, many incorrect matches may occur. Therefore, for completeness, the performance of feature extraction algorithms is computed using mAP score. To this end, first a Precision-Recall (P-R) curve is generated by varying the matching strategy parameter th from 0 to 1. 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 P-R curves are generated for each scene, we then compute the mAP scores by determining the area under the curves.

Chapter 3

Local Feature extraction in HDR Imagery under Drastic Lighting Variations

Drago Chiu Fattal LDR

Figure 3.1 – Harris corner detection in a lighting setup from Project Room (Row-1) and Light-Room (Row-2) datasets with local, global TMOs and best exposures LDR.

3.1 Overview

Adverse lightening conditions can significantly deteriorate the performance of keypoint detectors and descriptors in conventional LDR imagery. Several local and global normalization models [44, 103] have been designed to obtain better luminance invariant features. But these techniques are somewhat inefficient in practice. Poor performance of these algorithms

mainly accounts to the loss or change in the spatial configurations of the details present in a scene [110, 118].

HDR imagery brings potential to surpass these limitations and consequently enhance the feature extractor's output as accounted in [11, 19]; thanks to its wider dynamic range which enables to capture details in both dark and bright regions. However, it is not clear which are the best methods of employing HDR and weather these gains are significant on a real dataset.

This chapter investigates the potential of HDR for feature extraction stages *i.e.*keypoint detection and description, and in particular, addresses the following research questions:

1. *is HDR capable to achieve substantial quantitative gains in terms of feature stability to luminance changes compared to LDR*? are these gains consistent?

2. *which is the best way to use such HDR images*, i.e., direct real-valued luminance, or HDR converted to LDR format through a tone mapping operator (TMO) in order to be compatible with standard feature extraction techniques?

To answer these questions, an evaluation framework is provided in this chapter. Initially, we build a dataset of HDR and LDR images, consisting of two setups, each one illuminated with seven and eight different lighting conditions, respectively. The dataset is challenging in terms of texture reflectance of objects, presence of shadows and variety of illumination sources. For each lighting scene, we then consider a number of image encoding formats, including linear or perceptually encoded HDR values, the subjectively best LDR exposure, and several local or global tone-mapped pictures. Next, we detect features from each lighting scene, and we compute the standard repeatability of detected interest points in all the other illumination settings, in order to estimate the average feature stability. This is accomplished using two popular corner point (Harris [45]) and blob detectors (SURF [8]).

Some previous frameworks for evaluating the detectors and the descriptors have been proposed in the literature as discussed in Section 2. [55] and [20] framework report an increase in the number of detected feature points using HDR based modalities over LDR. However, the number of detected feature points is not itself a sufficient indicator of detection performance. Additionally, based on their results, it is difficult to draw precise conclusions on what makes certain HDR modalities perform better than others. Pribyl et al. [11] presents an evaluation of the repeatability on simple images under different transformations (lighting, viewpoint, distance) for a few LDR/HDR modalities but *not* the original HDR values. Aforementioned studies, mainly focused on evaluating detector performance only, with one or more different HDR representations [12, 82]. [19] evaluated the full feature extraction pipeline but with only a single HDR representation.

In this chapter, we focus on standard measures of feature stability under illumination changes along with analyzing the performance of many popular tone-mapping approaches which have been evaluated thoroughly from a perceptual point of view, but whose effectiveness in feature extraction has not been investigated so far. Additionally, we explicitly

compare direct feature extraction stages on HDR images with a tonemap-then-extract approach [55].

In a nutshell, three major contributions in this chapter are:

1. HDR Luminance Change Dataset,
2. Evaluation of Keypoint Detectors in HDR imagery,
3. Evaluation of full features extraction pipelines in HDR imagery.

3.2 HDR Luminance Change Dataset

Accuracy measurement of feature detection and matching is based on RR criterion and mAP, which relies on the precise localization of key-points in both reference and test images, so that correspondences between detected features points can be unequivocally assessed [96]. Unfortunately, the great majority of existing HDR image and video datasets are not adapted to this end, as images are not geometrically calibrated. The only such existing HDR dataset adequate for a confined low-level evaluation has been proposed in [11] (we refer to it as *2D and 3D Lighting Dataset* in the rest of this chapter), where a scene with controlled lighting conditions has been captured. However, the number of lighting conditions is quite limited.

In this chapter, we propose two different lighting setups: *Project Room* and *Light Room* (Figure 3.1), focusing mainly on lighting changes and variation in dynamic range of the scenes, which are recognized to be some of the most critical points in LDR feature detection, and are those for which HDR technology could bring most benefits.

Project Room (PR). The setup is composed of 8 different lighting scenes created by blocking light coming from a projector with the help of different objects. For each case, images with varying exposure time were captured using a Nikon D3100 digital camera. The setup is composed of several bright and dark colored objects arranged so as to create sharp shadows and overexposures in detailed areas. Created shadows hide the minute details for, e.g., bottom prints on memento, web-cam box printings etc.

Light room (LR). The dataset is composed of 7 different natural lighting conditions built by changes in global lighting due to opening and closing of window blinds, room ambient illumination and a diffused lighting from a tungsten lamp. For each condition, 6 images with different exposure time were shot using a Canon EOS 600D. This setup is also composed of dark and light objects with different type of object surfaces.

Both datasets, as shown in 3.1, are calibrated to the true physical luminance using the Minolta LS-100 Luminance meter, and can be downloaded from `http://webpages.l2s.centralesupelec.fr/perso/giuseppe.valenzise/sw/\ac{hdr}%20Scenes.zip`. In Figure 3.2, the variation in *dynamic range* of each scenes is provided. Image key [3] takes values on $[0, 1]$ and gives a measure of the overall brightness of the scene. Dynamic range

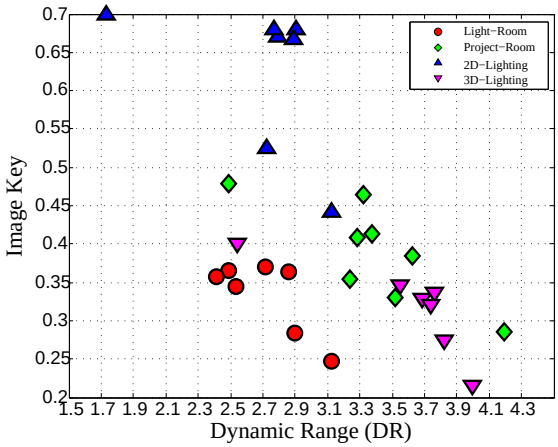


Figure 3.2 – Dynamic Range Vs Image-Key plots for (a) Light-Room(LR). (b) Project-Room. (c) Lighting 2D. (d) Lighting 3D [11].

is defined as $\log_{10}(L_{\max}/L_{\min})$, where $L_{\min}$ and $L_{\max}$ are the minimum and maximum HDR brightness values, respectively. Both properties give an indication of the variety of illumination conditions contained in the dataset.

3.3 Evaluation of Keypoint Detectors in HDR imagery

3.3.1 Keypoint Detectors

Feature extraction has been studied in vast details in computer vision literature where several techniques have been proposed and evaluated as detailed in Section 2.2, taking into account different challenging transformations. In this section, we focus on the two most widely used interest point detection schemes, i.e., *corner* and *blob* detectors, which are often used in several real time applications. In spite of several existing schemes for these approaches, we select two common detectors that have been used in similar evaluations for LDR content [12, 20].

For corner interest point detector, we employ the popular *Harris corner point* (Harris) detector [45], which is based on the autocorrelation score computed from local intensity change in an image. For blob detection, the experiments are carried out with the highly robust *SURF*[8] detector.

3.3.2 Considered LDR/HDR modalities

For each illumination change dataset, we consider the following LDR and HDR modalities:

- **LDR** best exposed image: we take the subjectively best LDR exposure shot for each illumination setup, i.e., the one that a human surveillance operator would select based on large details with smallest area of over- or under-exposed pixels;

Abbreviations	Description	L/G
D | Drago	An Adaptive logarithmic mapping [27]	G
W | Ward	Mapping based on histogram adjustment [58]	G
A | Ashikhmin	Gradient based mapping algorithm [4]	L
C | Chiu	Spatially non-uniform scaling algorithm [21]	L
M | Mantiuk	Perceptual method for contrast processing [67]	L
F | Fattal	Gradient domain HDR compression [35]	L
P | Pattnaik	Adaptive gain control for HDR [79]	L
R | Reinhard	Photographic tone reproduction method [90]	L
S | Schlick	Quantization techniques for visualization[95]	L

Table 3.1 – Local(L) and Global(G) TMOs.

- Tone-mapped image: we consider two global **(GTM)** and seven local **(LTM)** TMO's (see Table 3.1) to convert HDR pictures to 8-bit LDR, which are representative of the most popular tone mapping techniques for rendering HDR on LDR displays proposed in the literature [13];
- HDR linear values **(HDR-Lin)**, i.e., photometric luminance values stored in the HDR file;
- HDR perceptually encoded values: we consider a simple logarithmic **(HDR-Log)** encoding, according to Weber-Fechner's law; or the perceptually uniform encoding **(HDR-PU)** proposed in [5], which accounts for the drop of sensitivity at lower luminance levels. Notice that PU encoding needs photometrically calibrated HDR pixels as input. Both Log and PU values are rescaled in $[0, 1]$.

In total, 13 different image formats are thus considered for each lighting condition. We stress the difference between HDR encoded values and GTM pixel values: the former are the result of a simple transfer function and encoded using floating point values; the latter, instead, are the result of a content-dependent operation, and are encoded on 8-bit, integer precision.

3.3.3 Experimental Results and Discussion

Experiments are carried out on proposed datasets (LR,PR) and lighting dataset (2D,3D) of [11]. The only measure of accuracy considered in this chapter is the repeatability rate (R-score) as discussed in Section 2.2. For this evaluation, we used $\epsilon = 35$px, which is less than 1% the image size, similar to [11]. Also the evaluation scheme is confined to the strongest 200 key-points in marked RoI's (Region of Interests). This not only limits the feature point detection in pertinent areas, but also helps to ensure a fair comparison of the blob or corner key-point detection on diverse datasets, as different detectors result in varying number of keypoints which can bias the R-score.

The experimental study is conducted in two phases. In the first phase, for each dataset, one scene is selected as a reference image, and the repeatability is computed with the other scenes (test images). Relative gains are recorded for all best HDR based modalities (GTM,

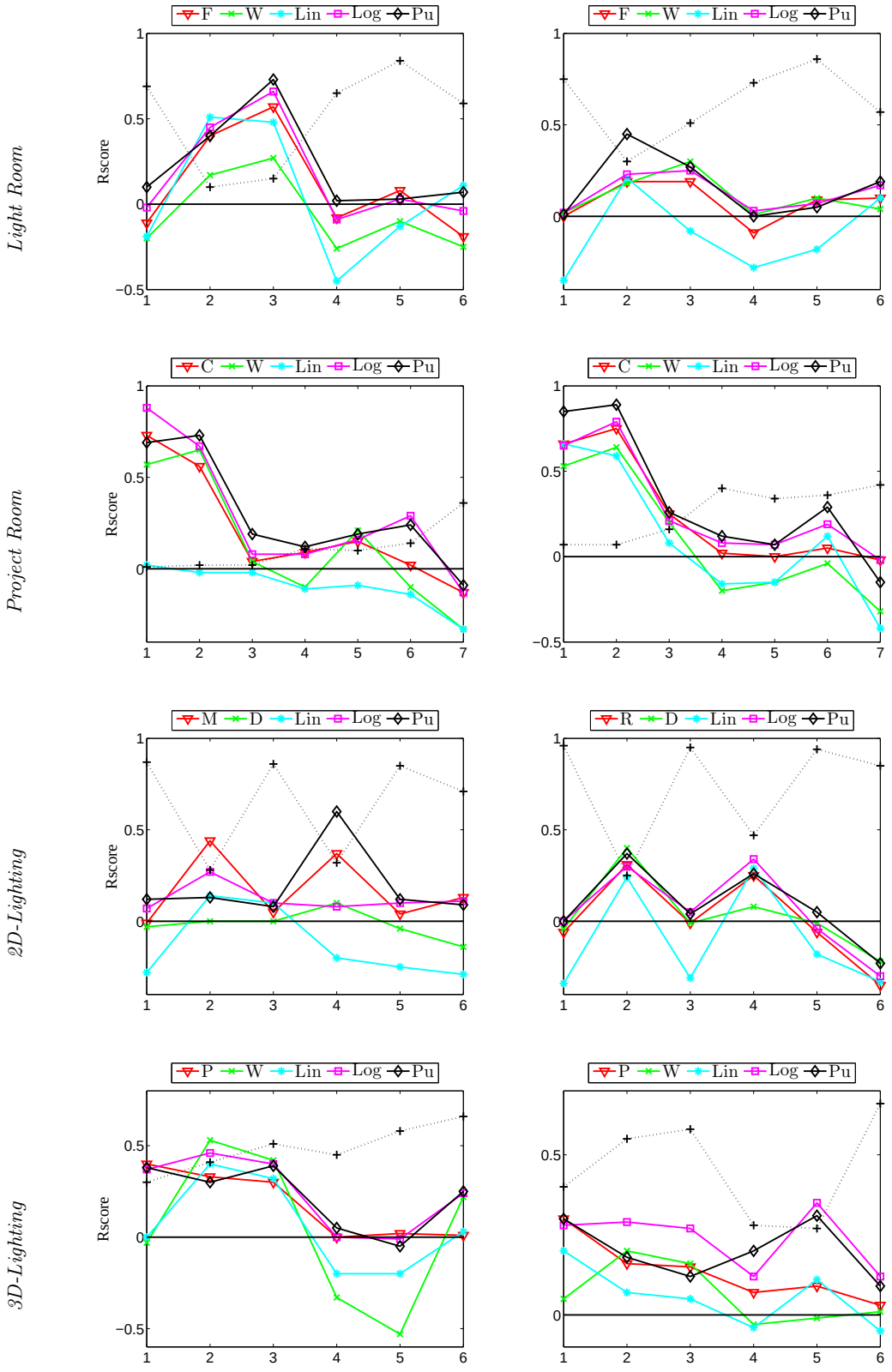


Figure 3.3 – Relative gains by best LTM, GTM (abbreviated using Table 3.1), Linear, Log and Pu HDR encodings with respect to LDR for different test datasets (scenes indicated by progressive numbers on x-axis). The dotted line shows the absolute R-scores of LDR.

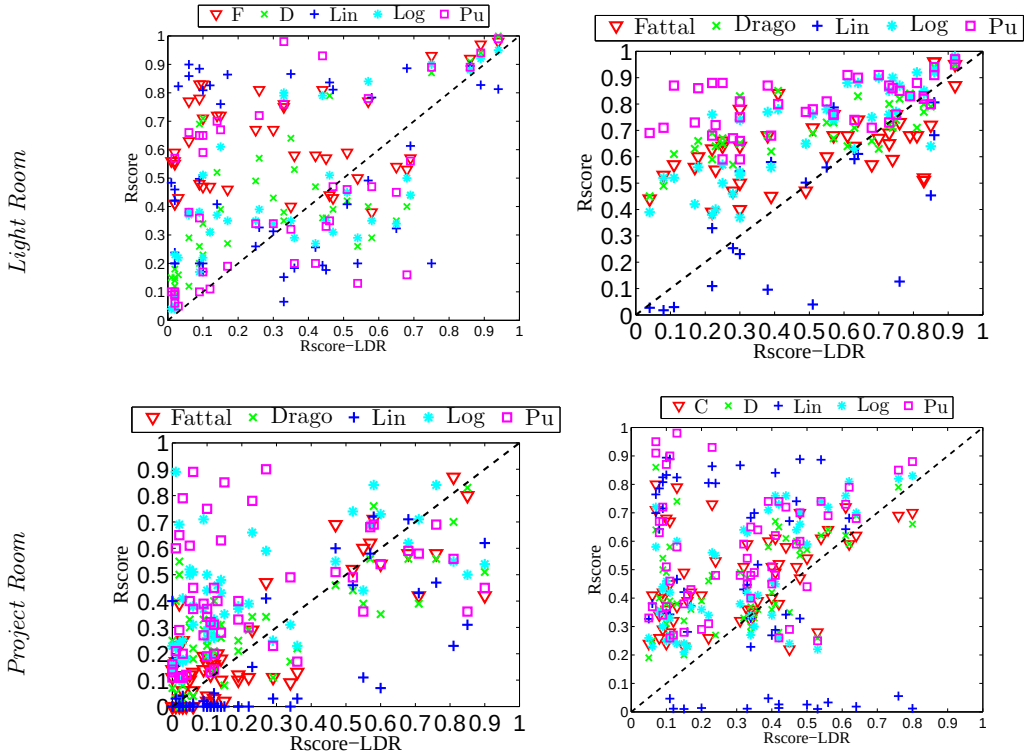


Figure 3.4 – Scatter plots for HDR based formats (TMOs abbreviated using Table 3.1) with respect to LDR on proposed dataset using detector: Harris and SURF

LTM, and HDR encoded formats) with respect to LDR, by subtracting the LDR R-score from each individual format as shown in Figure 3.3. The black dotted line depicts the absolute LDR R-score for each test image pair. For each dataset and using either detector, we observe high relative gains by HDR based modalities (especially by HDR-PU), but still, they are not positive everywhere (e.g., scene 0(ref) – 7 of Project Room dataset and scene 0(ref) – 6 of 2D-Lighting dataset test pairs from Figure 3.3).

In the second phase of the experiments, in order to determine more concrete quantitative information about all such possible cases, we expand our experimental test bench by involving all the possible images pairs for both LR and PR dataset, i.e., each condition is in turn the reference and the others are the test images. In this phase, we firstly determine the relative performance of the best performing HDR based formats, i.e., encoded HDR, LTMs and GTMs, with respect to the traditional LDR, producing the scatter plots shown in Figure 3.4 and 3.5. Each scatter plot shows the relative gain with respect to a compared format, while the dashed line is the 45° line: points lying above this line shows higher performance and points lying below are performing lower than the compared format. The distribution of points of tone mapped and encoded HDR based formats above the line, implies that these formats are capable to capture wider range of information from the images than the respective LDR format as shown in Figure 3.4. However, this is not true for HDR-linear format which shows the worst overall performance. On the other side, we also

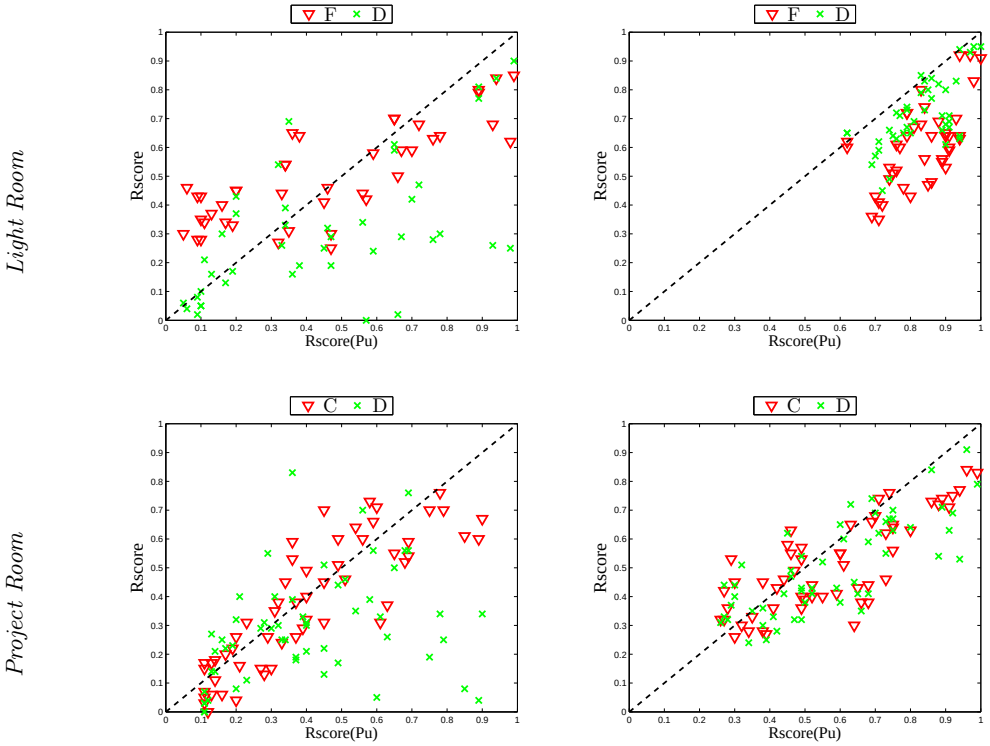


Figure 3.5 – Scatter plots for best performing GTM and LTM (abbreviated using Table 3.1) with respect to HDR-Pu encoding on proposed dataset using detector: Harris and SURF

investigate the relative performance of the tone mapping and best HDR encoded format in Figure 3.5. The results obtained suggests that in many cases, applying a TMO entails a loss of detected keypoints.

In addition, the averages and standard deviations of the gains in R-score of all HDR modalities over LDR are shown in Figure 3.6. These are obtained by subtracting the R-score of the individual format from the absolute LDR R-score. In the following, we comment on the performance of the different HDR and LDR modalities for feature point detection.

HDR versus LDR. In all conditions and for both key-point detectors, average values show significant gains of HDR or tone-mapped images over single LDR exposure. This is consistent with what has been found in [19]. However, based on the results from scatter plots in Figure 3.4, we observe that there are some scenarios where best LDR records higher performance than the rest of the HDR based formats. We believe that this is mainly due to significant illumination differences in pertinent regions of image pairs.

HDR encodings. The best average repeatability scores are in general obtained with PU-HDR encoded values. This is not surprising, as HDR formats store most of the pertinent information in the scene, and it is therefore promising to research towards application of feature extraction on these modalities. From results, it is also clear that these encodings give significantly better results than photometric HDR-Lin. This is a non-obvious conclusion of this work, i.e., that HDR-Lin is not appropriate to be used for feature extraction

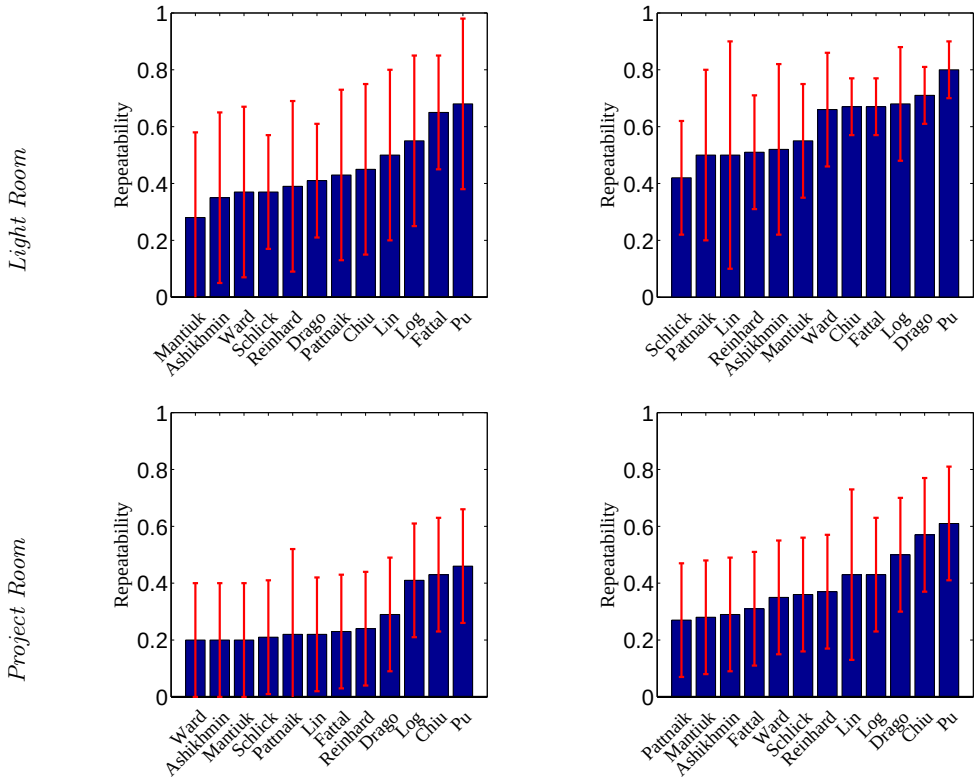


Figure 3.6 – Average gain recorded by different formats over the LDR.

algorithms, especially by observing huge variation in its behavior. This seems to suggest that feature extraction algorithms designed and optimized for LDR content somehow require the optimally scaled pixel values. Also for such algorithms, even the parameter tuning is not an option as the range of photometric luminance value can largely vary with the content.

HDR versus TMO's. Average R-scores gains over LDR by tone mapping techniques are either comparable or lower than those of HDR encoding formats. However better performances recorded by some TMO's draws significant attention, such as Drago [27], Chiu [21] and Fattal [35]. In addition to the performance evaluation for LTMs', it is interesting that the gradient-based local techniques, i.e., Fattal and Chiu TMO's, have shown comparable gains than other LTM techniques, in specific scenarios. This is inverse to observations in perceptual applications [13], where these two LTMs' are deemed as worst performers. This further establishes that there is less congruency between visually pleasing tone mappings and vision-task-based optimal mapping technique. Another important point to note here is that these tone mappings perform better with blob detectors than corner point, and it is consistent with the observations in the literature comparing detectors [96]. In addition to all the observations, it is also worth mentioning that there is no unanimous winner amongst these tone mapping techniques using either detection criterion.

Abbreviations	Description
LDR	Best exposure LDR image of the scene
RNG(G)	A global scaled mapping operator [90]
DR(G)	An Adaptive logarithmic mapping [27]
RN(L)	A local dodging-and-burning operator [90]
MA(L)	Perceptual method for contrast processing [67]
FA(L)	Gradient domain HDR compression [35]
CH(L)	Spatially non-uniform scaling algorithm [21]
DU(L)	A fast bilateral filtering technique [29]
HDRLog	Logarithmic encoding in accordance to Weber-Fechner's law
HDRLin	Linear photometric luminance values stored in the HDR file

Table 3.2 – Different image modalities for feature extraction.

3.4 Evaluation of full Features Extraction Pipelines in HDR imagery

3.4.1 Considered LDR/HDR modalities

We consider a total of 10 different image modalities (listed in Table 3.2) including the standard 8-bit LDR, 2 floating point HDR representations (HDRlog and HDRlin) and 7 different 8-bit TMO HDR representations. These consist of 2 global and 5 local TMOs. In this chapter, we have considered a subset of the TMOs introduced in Table 3.1. The choice has been entirely based on trade-off between their performance in keypoint detection task in Chapter 3 and perception based tasks in [13]. Additionally, note that we do not include PU encoded HDR. This is mainly due to the absence of luminance-based calibrated scenes from 2D and 3D lighting dataset.

3.4.2 Feature extraction

In this section, following 4 popular feature extraction schemes are assessed. Both gradient-based histograms and computationally fast binary descriptors are employed for the evaluation.

- **SIFT** [64]. This classic scheme is constituted of a blob keypoint detector (based on difference of Gaussians) and a gradient-based descriptor. The SIFT descriptor is a 128-dimensional histogram formed by concatenation of the image gradients computed on 4x4 grid spatial neighborhood around the detected keypoint.
- **SURF** [8]. SURF scheme is composed of a computationally efficient blob type detector mainly based on the Hessian matrix approximation along with a descriptor computed as the sum of the Haar wavelet response around the point of interest.
- **BRISK** [61]. With major focus on computational efficiency, the BRISK feature extraction is made up of a fast multi-scale detector and a binary descriptor. The detection module is an extension of corner-based detectors like AGAST and FAST. The descriptor is a binary string computed by brightness comparisons on circular sampling patterns around the detected regions.

Project Room | Light Room | 2D Lighting | 3D Lighting

Figure 3.7 – Example images from the datasets.

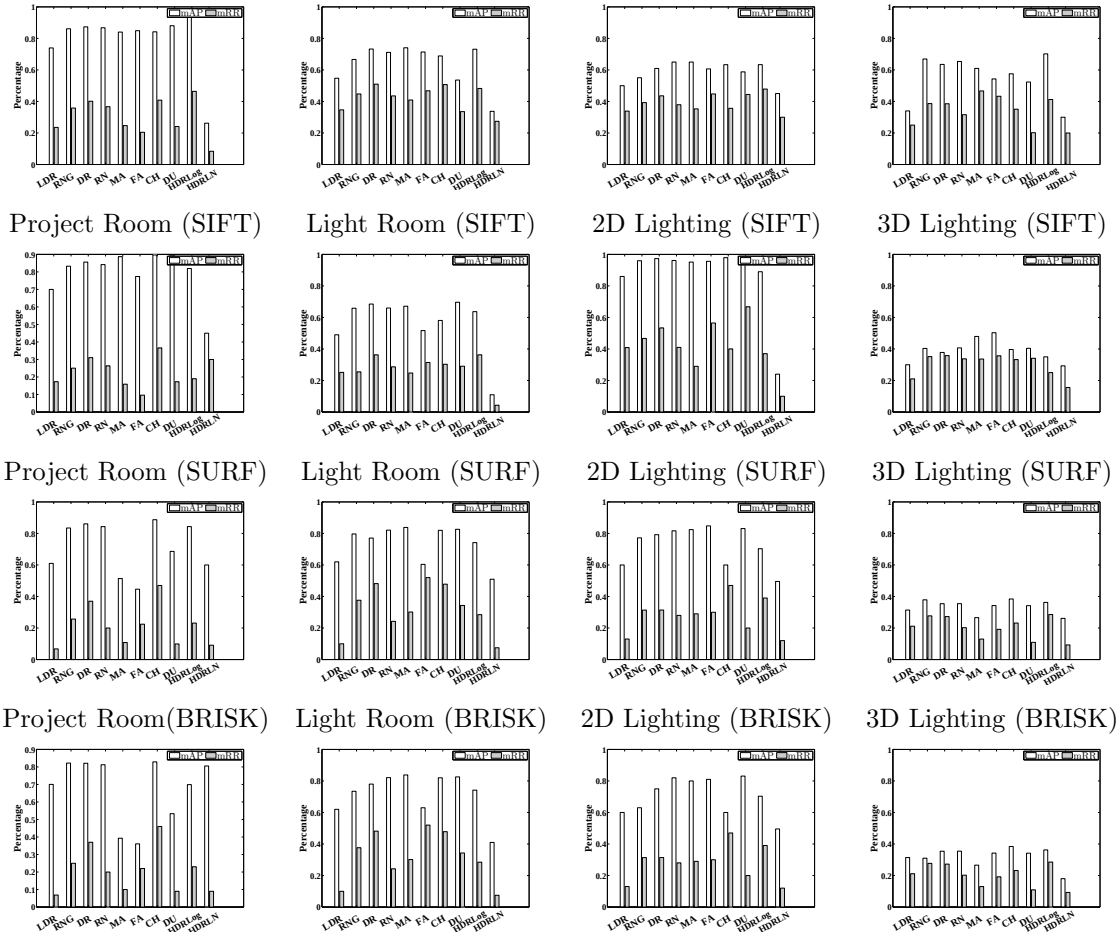
- **FREAK** [77]. Similar to BRISK scheme, it has the same BRISK detector along with a binary descriptor called FREAK. Similar to BRISK descriptor, FREAK also uses a concentric rings arrangement, but the sampling grid is non uniform as inner circular rings have exponentially more points.

3.4.3 Experimental Results and Discussion

Our test setup comprises a total of 29 images (8 Project Room + 7 Light Room + 7 2D-Lighting + 7 3D-Lighting) for each image representation. In the first part of experimental validation, we look at the overall feature extraction performance, by computing the mAP over all datasets. To this end, we evaluate matching using a test bench of 182 image pairs (56 Project Room + 42 Light Room + 42 2D-Lighting + 42 3D-Lighting). Following the detection protocol from [11, 82], we first select 400 keypoints with the strongest detector response and measure the keypoint stability using the RR. Then, for the descriptor part, we compute standard precision-recall (P-R) curves [70] for measuring the accuracy of matching. For each image pair, we compute the PR curve by varying th from 0.0 to 1.0 and record the average precision value. After this, for each format and either feature extraction scheme, a mAP score is obtained by averaging the average precision calculated on such 182 image pairs (see Table 3.3). Both metrics RR and mAP are defined in Section 2.2.3.

Furthermore, to understand how detector and descriptor contribute to the overall performance, we expand our analysis to individual datasets and compute mAP and RR. In Figure 3.8, we report side-by-side the mAP and RR for each extraction scheme for all datasets, respectively. It is evident that in most of the cases higher RR entails higher mAP scores, i.e., having more stable keypoints strongly influences the overall matching performance. Nevertheless, there are few exceptions, e.g. RN and FA in 3D-Lighting dataset, discussed later in this Section. In the following, we examine in detail the main conclusions obtained from our results.

HDRLin versus all. The results in Table 3.3 show that HDRLin representation is consistently the worst performing using all extraction schemes. This is coherent with the



Project Room (SIFT) Light Room (SIFT) 2D Lighting (SIFT) 3D Lighting (SIFT)

Project Room (SURF) Light Room (SURF) 2D Lighting (SURF) 3D Lighting (SURF)

Project Room(BRISK) Light Room (BRISK) 2D Lighting (BRISK) 3D Lighting (BRISK)

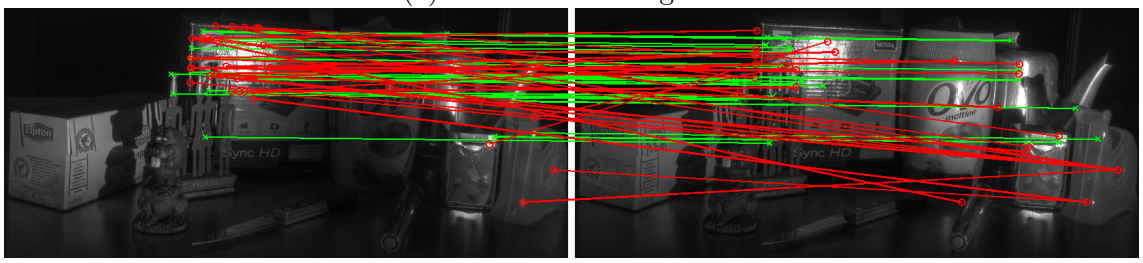
Project Room(FREAK) Light Room (FREAK) 2D Lighting (FREAK) 3D Lighting (FREAK)

Figure 3.8 – Mean average precision (mAP) and mean repeatability rate (mRR) over the four considered datasets and feature schemes. mAP and mRR are computed on 56 image pairs, for the Project Room dataset, and over 42 image pairs for Light Room, 2D and 3D Lighting datasets.

Repr.	Feature Extraction Schemes				**Avg/Repr.**
	SIFT	SURF	BRISK	FREAK	
LDR	55	62	60	61	59.5
RNG	69	70	71	65	67.5
DR	72	72	71	**73**	<u>72</u>
RN	72	70	**73**	72	<u>72</u>
MA	74	**75**	62	62	68.3
FA	68	67	62	66	65.8
CH	68	71	64	66	67.3
DU	64	72	68	71	68.8
HDRLog	**75**	66	67	68	69
HDRLin	44	30	50	41	41.5
Avg/Schemes	<u>66.8</u>	65.6	65.5	65	

Table 3.3 – Mean Average Precision (mAP %) scores for the 10 considered representations using 4 feature extraction schemes. Scores are averaged over 4 lighting change datasets. Highest mAP score for each scheme is shown in **bold**. Best Avg/Formats and Avg/Schemes scores are <u>double underlined</u>.

(a) Feature Matching - FA



(b) Feature Matching - RN

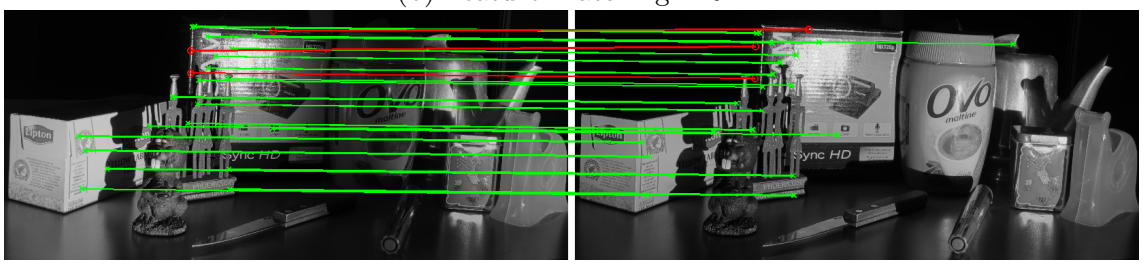


Figure 3.9 – An example of image matching for two TMOs. The true positive and false positive matches are shown with green and red lines respectively. The TM in (a) achieves a higher repeatability (24 %) than that in (b); however, most of the matches in (a) are false positives, thus the AP for (b) is higher than in (a) (95 % vs. 87 %, respectively).

previous findings in [12, 82], and is mainly due to the low keypoint repeatability, which increases the probability of false positives. This leads to the first conclusion of this work, i.e., HDRLin is not appropriate to be used for feature extraction algorithms, for both detector and descriptor.

HDRLog/TMO versus LDR. On average, all HDR formats show significant gains of (at least) 8% mAP over single LDR exposure (see Avg/Formats in Table 3.3). This partially accounts to having more false matches in LDR due to loss in local textural information in lighting transformations. Another reason which is evident from Figure 3.8, is the low repeatability rate which reduces the number of true positives.

HDRLog versus TMO's. mAP scores obtained from HDRLog and different TMOs are relatively comparable. This implies that there are not significant advantages in using a floating-point HDR representation over 8-bit TMs. Alternative HDR encodings could improve further mAP scores, such as the PU encoding [5], as reported for keypoint repeatability in [82]. However, those representations require photometrically calibrated HDR pictures, which might not be available in practice.

Comparison with previous studies. Previous studies [12, 82] have reported that local TMO approaches such as Fattal or Chiu consistently provide more stable keypoints (in terms of repeatability) under illumination changes, compared to TMOs which are generally considered good from a perceptual perspective, such as Reinhard. The results of this work show that those trends are less evident when the overall feature extraction pipeline is considered. For instance, from Figure 3.8 we observe that some TMOs achieve better repeatability rates but lower overall mAP scores compared to others formats, e.g., this is the case for RN and FA tone mappings in Project Room and Light Room dataset using BRISK and FREAK, or for RN and FA in 3D Lighting dataset using SIFT. We deduce that in those cases, although the fraction of repeated keypoints is lower, the corresponding descriptors are more discriminative, i.e., they yield a lower rate of false positives, or equivalently, a higher portion of matches are true. Figure 3.9 shows an example of image matching for the Project room dataset, using RN and FA tone mappings and BRISK features. It is clear that, although the number of matches is lower in RN, they are "better quality", in the sense that most of them are true positives. Conversely, in FA, although the basis of possible matches is larger, most matches are indeed false, which reduces the average precision as reported by the mAP scores in Figure 3.8.

Another important point to note is that these tone mappings perform well with all feature extraction scheme for different lighting transformations, with marginal gains for SIFT. In addition to all the observations, it is also worth mentioning that there is no unanimous winner amongst these tone mapping techniques for all extraction criterion.

3.5 Conclusion

In this chapter, we presented a comprehensive evaluation of different HDR and LDR based modalities for visual feature detection and matching, under changes in illumination conditions. The analysis based on the RR and mAP scores on different scenes confirms the potential of HDR techniques over single LDR exposures. For both detection and matching, our results confirms that the linear HDR values are inappropriate to be used for visual recognition tasks. Furthermore, we observe that local TMO's producing very appealing results in terms of rendering quality are not necessarily the best option for image analysis. More interestingly, we have also observed that local TMOs with very high repeatability rate for feature detection are not necessarily the best option when the full feature extraction pipeline is considered. Although we measured a consistent gain in the average repeatability scores when using direct HDR pixel values over tone-mapping, it does not hold true for the test pairs individually. Hence, it is difficult to comment what is better before extracting the features, a) to encode the HDR pixel approximatively linear to perception, or b) directly tone map.

This study further strengthens our argument that there might be quite a large room for improvement in feature extraction performance at detection and description stages by designing optimal tone mapping schemes for HDR, which can ensure high average precision as well as repeatability rates, and that can be easily fused with current recognition algorithms.

Consequently, in the following chapter, we will primarily investigate the key criteria for designing the optimal TMO, starting with the keypoint detection. Then, we will further approach to designing a corresponding optimal TMO which aims at enhancing the efficiency in keypoint detection.

The work presented in this chapter has resulted in the following publications:

1. A. Rana and G. Valenzise and F. Dufaux, "Evaluation of Feature Detection in HDR Based Imaging Under Changes in Illumination Conditions", *IEEE International Symposium on Multimedia (ISM)*, Miami,USA, December, 2015.

2. A. Rana and G. Valenzise and F. Dufaux, "An Evaluation of HDR Image Matching under Extreme Illumination Changes", *The International Conference on Visual Communications and Image Processing (VCIP)*, Chengdu, China, 2015.

Chapter 4

Tone Mapping Operator for Efficient Keypoint Detection

4.1 Overview

TMOs have traditionally been designed to display HDR pictures in a perceptually favorable way and mainly preserve the human-vision attributes such as image aesthetics and perceptual contrast. However, when such tone-mapped images are to be used for computer vision tasks such as *keypoint detection*, these design approaches are suboptimal [11, 82, 83] and needs to be re-calibrated. No related work exists in the literature, which aims at designing a detection-optimized tone mapping technique or comprehending the related criteria involved.

In this chapter, we address the problem of optimal TMO design for keypoint detection task. Specifically, we investigate the following questions a) what are the factors to be considered in the TMO design when targeting keypoint detection tasks ?, and b) how can we optimize a TMO for such tasks under drastic illumination variations.

To answer the aforementioned questions, this chapter initially discusses the suboptimality of existing TMOs and derives guidelines to design a keypoint-optimized TMO. To that end, a comparison is drawn between the optimization of existing TMO parameters with respect to: a) Repeatability Rate RR and b) correlation coefficient CC between tone-mapped images of the same scene with lighting variations. CC measures the statistical similarity between a pair of tone-mapped images. The goal here is to find whether optimizing a TMO with respect to RR leads to higher keypoint stability over the per-pixel similarity (using CC) between the tone-mapped images.

Building upon the observations from optimality study, in this chapter, we introduce a novel learning based adaptive TMO for robust keypoint detection. Our proposed framework aims at enhancing the repeated detection of sparse keypoint locations (*e.g.* corners) in high-contrast areas of scenes undergoing complex real-world illumination transitions such as day/night change. To this end, we initially introduce an adaptive TMO which can be

locally modulated, *i.e.* its parameters can vary pixel-wise. Then, the per pixel modulation is derived by means of a learned illumination invariant model. In this context, we train a Support Vector Regressor (SVR) to predict the desired pixel-wise modulation maps by using the linear HDR content from scenes captured with varying lighting conditions.

Learning-based models have been seldom pursued for designing keypoint-optimized TMOs. As a consequence, there is no standard dataset to train or test any model in this context. In this chapter, we overcome this difficulty by proposing a simple detection-similarity-maximization model to generate appropriate training samples. Additionally, we propose an HDR dataset of 8 image scenes taken in indoor and outdoor locations with different lighting variations.

In nutshell, there are three major contributions in this chapter,

1. Factor for optimizing a TMO for Keypoint detection,
2. Learning based adaptive tone mapping model for efficient keypoint detection,
3. A Luminance change HDR dataset.

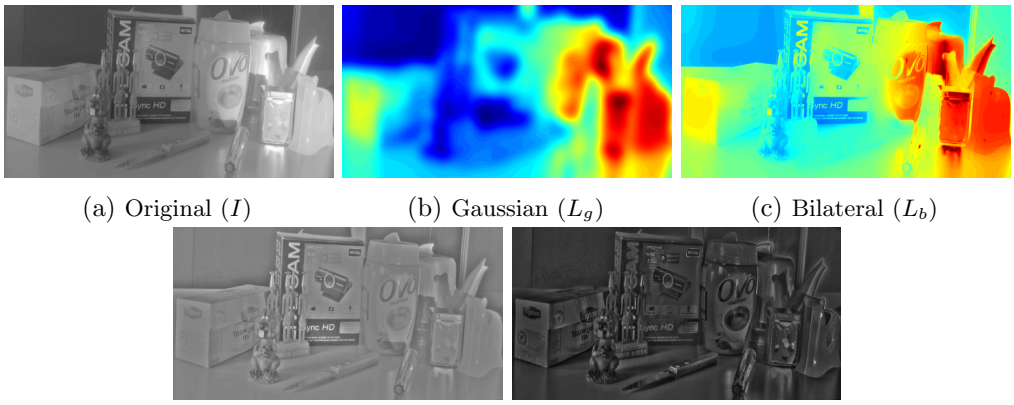
4.2 Optimizing a TMO for Keypoint detection

In this section, we present a study on the parametric optimization of TMOs using two factors: CC and RR, aiming at enhancing the keypoint detection performance under drastic lighting change scenarios. Our main aim is to find out what factor could be the most interesting to define an optimized TMO.

RR is a conventional performance measure of keypoint detection algorithms and is computed on repeated occurrences of detected keypoints in test and reference images. Hence, optimization of TMOs with respect to RR could estimate the optimal detection performance gains.

On the other hand, the CC computes the statistical similarity between distribution of images. Theoretically, a CC-optimized TMO should improve keypoint detection performance. This is mainly because as a high statistical similarity of tone mapped images should increase the probability of detection of similar keypoints. Such scenarios in return could be highly interesting for optimizing a potential class of TMOs which are based on illumination normalization such as [21, 29]. In simple words, what if a substantial improvement in keypoint detection performance can be achieved just by estimating the ideal reflectance maps from the HDR images using TMOs such as [21, 29].

In the following, we first detail the considered TMO. Next, we briefly discuss the feature detection methods, followed by metrics and dataset selection. Finally, we describe the optimization strategies of considered models.



(a) Original (I) (b) Gaussian (L_g) (c) Bilateral (L_b)

(d) Gaussian Reflectance(R_g) (e) Bilateral Reflectance(R_b)

Figure 4.1 – Reflectance images R_g and R_b from original image I using the Gaussian and Bilateral luminance maps L_g and L_b respectively.

4.2.1 Considered TMO

In this study, we employ two well-known Retinex inspired approaches using: a) Gaussian model [21] and b) Bilateral model [104] for tone mapping. We choose these TMOs as they are promising to bring illumination invariant (as much as possible) reflectance maps; an ideal case for efficient keypoint detection.

According to Retinex theory of physical image modeling, we assume that I, the HDR image to be tone mapped, is the product of the luminance L of the scene (which varies with different illumination conditions) and of the reflectance R characterizing objects of the scene, i.e., $I = R \cdot L$. The luminance L is generally assumed to be spatially smooth [42], while reflectance contains fine-grained details, texture and edges which are relevant for detection [70, 96]. Once L is estimated, the final reflectance image is given by $R = I/L$.

In the following, we briefly describe the aforementioned luminance estimation TMO:

a). **Gaussian tone mapping** (GTM) model gives the reflectance image R_g as $R_g = I/L_g$ where

$$L_g = I * G_\sigma, \tag{4.1}$$

where G_σ is a Gaussian filter where the parameter σ depends on image size $[m \times n]$, i.e. $\sigma = \alpha \cdot \max(m, n)$. When targeting visual perception, the parameter σ is tuned so as to reduce visual artifacts like halos observed around detected edges. This model with a ingle parameter is simple and computationally very fast.

b). **Bilateral tone mapping** (BTM) model is a a precise, non-linear and edge preserving filter where R_b is computed as $R_b = I/L_b$. The luminance estimation $L_b(x)$ is given as:

$$L_b(x) = \frac{1}{N} \sum_{y \in S} G_{\sigma_s}(\|x - y\|) \cdot G_{\sigma_r}(\|I_x - I_y\|) I_y, \tag{4.2}$$

where x and y are pixel locations, S is the set of neighborhood locations, G_{σ_r} and G_{σ_s}

are Gaussian filters with variances σ_r and σ_s referred to as range and spatial parameter respectively. N is a normalization factor term :

$$N = \sum_{y \in S} G_{\sigma_s}(\|x - y\|) G_{\sigma_r}(\|I_x - I_y\|). \tag{4.3}$$

It is important to note that when the range parameter increases, the model gradually approaches Gaussian convolution. This is mainly because the Gaussian G_{σ_r} widens and flattens, and essentially, it becomes nearly constant over the intensity interval of the image. Conversely, when the spatial parameter increases, larger details like edges get smoothened in the image.

An example of luminance estimation with their corresponding reflectance image is shown in Figure 4.1.

4.2.2 Keypoint point detection

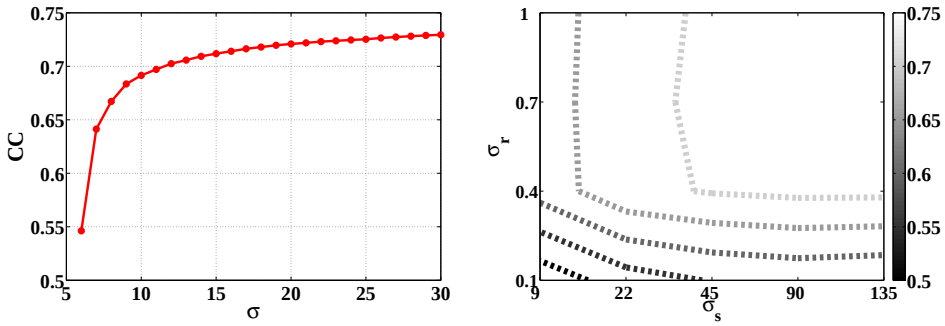
Keypoint detection has been widely studied in computer vision literature where several techniques have been proposed and evaluated [96] taking into account different challenging transformations. In this study, we focus on the two most widely used keypoint detection schemes, i.e., *corner* and *blob* detectors. We select two common detectors that have been used in previous HDR imagery based evaluations [11, 82] and are often used in several real-time applications. For corner interest point detector, we employ the popular *Harris corner point* detector [45], which is based on the autocorrelation score computed from local intensity change in an image. For blob detection, we carried out experiments with the highly robust *SURF*[8] detector.

4.2.3 Metrics

In this study, we build our framework using following metrics.

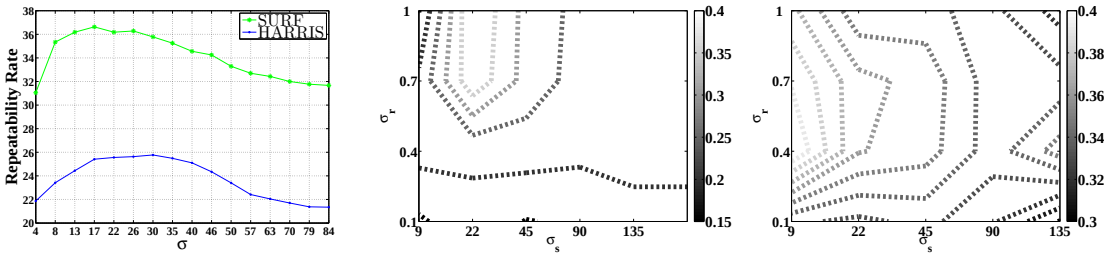
- RR (see Section 2.2) is a standardized method detailed in [96] to measure the detector accuracy. It is given as the fraction of keypoints detected in the reference image which are repeated in the test image to the minimum of a total number of detected points in test or reference image. A keypoint is considered to be repeated in the test image if: a) it is detected as a keypoint in the test image, and b) it lies in a circle of radius ϵ centered on the projection of the reference keypoint onto the test image. ϵ determines the keypoint detection error rate. RR is given as $\frac{R_i(\epsilon)}{\min(n_r, n_i)}$, where $R_i(\epsilon)$ is the number of keypoints detected in the reference image which are repeated in the test image, n_r, n_i is the number of detected keypoints in reference and test image respectively.

- *Correlation Coefficient* (CC) is well-known to quantify the strength of a linear relationship between two variables. In this study, we have used this metric to measure the correlation between two image maps. Values close to 1 indicate that there is a



(a) CC-GTM (b) CC-GTM

Figure 4.2 – *Parameters vs Correlation Coefficient (CC) for Project Room dataset.* (a) σ vs CC for Gaussian tone mapping (GTM) model. (b) σ_r and σ_s contours for Bilateral tone mapping (BTM) model with color magnitudes showing average CC scores.



(a) (b) (c)

Figure 4.3 – *Parameters vs Repeatability Rate (RR) for Project Room dataset.* (a) σ vs RR for both SURF and Harris detector for repeatability rate Gaussian TMO (RRGTM). (b) σ_r and σ_s contours for Harris detector and (c) for SURF detector for repeatability rate Bilateral TMO (RRBTM) with color magnitudes showing RR scores.

positive linear relationship whereas 0 points no linear relationship between image maps.

4.2.4 Datasets

We considered the HDR dataset with substantial illumination changes proposed in the previous chapter [82]. It is composed of 2 parts: Project-Room with 8 lighting conditions and Light-Room with 7 lighting conditions.

4.2.5 Optimization of TMOs

Let $I_1(x)$ and $I_2(x)$ be two images of the same scene, illuminated by two different illumination maps $L_1(x)$ and $L_2(x)$. According to Retinex theory $I_1(x) = L_1(x) \cdot R_1(x)$ and $I_2(x) = L_2(x) \cdot R_2(x)$. An ideal Retinex algorithm estimates $L_1(x)$ and $L_2(x)$ such that two ideal reflectance maps are equal $R_1(x) = R_2(x)$. In such an ideal scenario, keypoint detection performance should be enhanced considerably as the keypoints in reflectance maps would be identical.

However, Retinex is a mathematically ill-posed problem [50]. In practice, it often implies that $R_1 \neq R_2$. Besides this, all perceptually optimized Retinex based models aim at finding the best luminance estimation for a given scene such as $I_1 = L_1 \cdot R_1$, rather than optimizing on aforementioned image pairs like (R_1, R_2) which is the classical way of measuring keypoint detection performance.

Therefore, we firstly investigate the maximization of the correlation between reflectance images pair. The main motivation is that highly correlated reflectance maps should result in detected keypoints that are alike, thereby, enhancing the keypoint repeatability. Alternatively, we also investigate the optimization of considered models with respect to repeatability rate which will help to analyze the maximum gains achievable with conventional TMO.

In summary, we optimize the considered Retinex based GTM and BTM models in two ways: 1) with respect to correlation of reflectance maps of image pairs and, 2) with respect to detector repeatability. For the first method, we iteratively optimize the GTM and BTM models parameters with respect to CC on both datasets using each detector. More specifically, we iteratively tune the parameter σ for GTM and σ_s, σ_r for BTM with the aim of maximizing the overall CC using each detector.

Correlation based parameter tuning for Project Room dataset in illustrated in Figure 4.2 (a). It depicts that for GTM, a higher σ, i.e. high variance Gaussian blur, minimizes the absolute differences between the reflectance image pairs. The same observation also holds for range and spatial parameters of BTM model as shown in Figure 4.2 (b). Thereafter, using these correlation based optimized models, we generate the tone mapped images as correlation-coefficient-Gaussian-tone-mapping (CCGTM) and correlation-coefficient-Bilateral-tone-mapping (CCBTM) for both datasets.

For the second method of optimization with respect to detector repeatability, we tune the GTM and BTM parameters with the aim of maximizing the overall repeatability rate (RR) of keypoints using both Harris and SURF detector. Corresponding results are shown in Figure 4.3 (a) for the Gaussian model, and in Figure 4.3 (b) and (c) for the Bilateral model. Similarly, we generate the tone mapped images, repeatability-rate-Gaussian-tone-mapping (RRGTM) and repeatability-rate-Bilateral-tone-mapping (RRBTM), for both datasets and using each detector.

4.2.6 Experimental Results and Discussion

We evaluate the CCGTM, CCBTM, RRGTM, RRBTM optimized tone mappings models for keypoint detection using Harris and SURF detectors in Figure 4.4(a),(b). Additionally, we compare 5 different local and global high performing TMOs [21, 27, 29, 67, 90] with our RRGTM and RRBTM tone mapping models as shown in Figure 4.4(c),(d).

For each TMO, we measure the overall keypoint detection accuracy using the RR performance metrics. Initially, we compute the individual RR using a particular detector

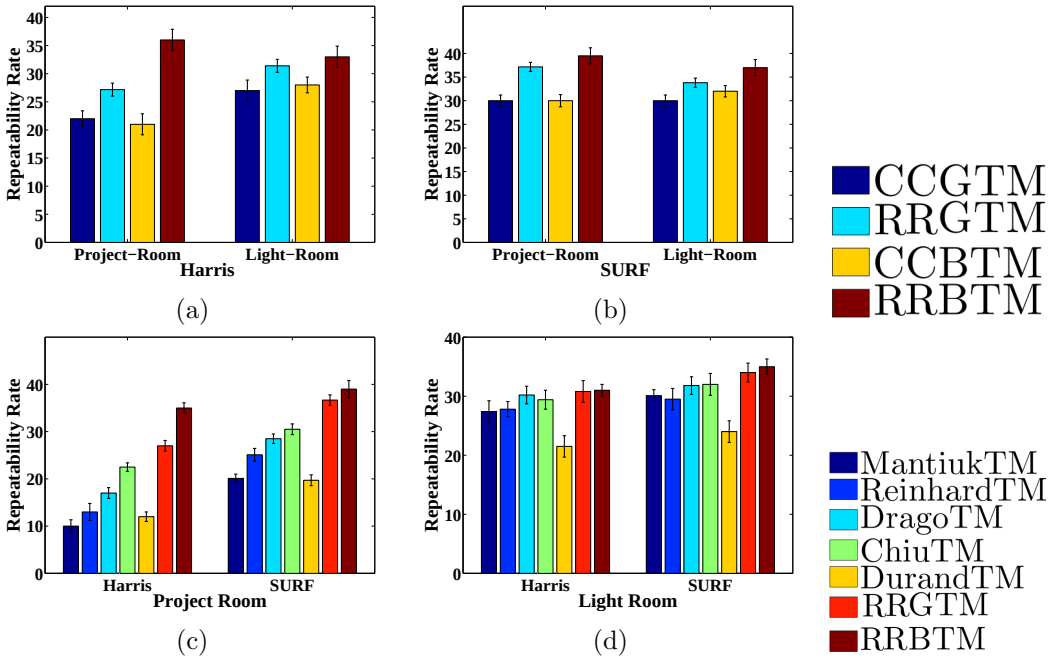
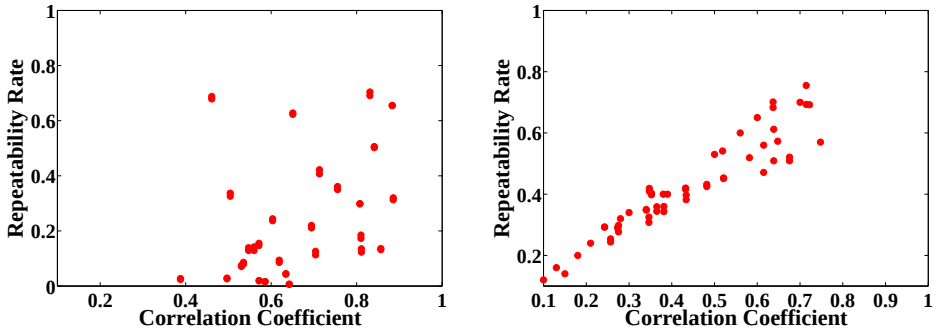
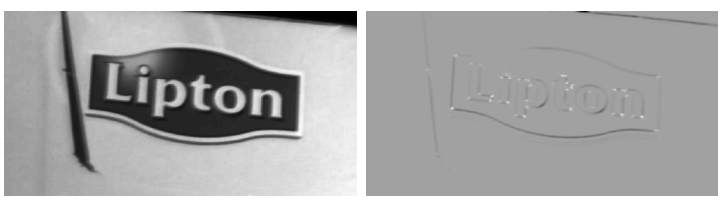


Figure 4.4 – Row 1. (a) and (b) Average repeatability score and standard deviation for the both correlation and response based optimized approaches using Harris and SURF detector respectively. Row 2. (c) and (d) Average repeatability score and standard deviation for the reflectance models (GTM and BTM) and other commonly used TMs on Project Room and Light Room dataset



(a) Reflectance map (b) Response Map

Figure 4.5 – *Scatter plots.* (a) correlation coefficients of reflectance maps $CC(R_i, R_j)$ vs corresponding repeatability rate $RR(R_i, R_j)$, (b) correlation coefficients of response maps $CC(Resp_m, Resp_n)$ vs corresponding repeatability rate $RR(R_m, R_n)$ for HDR log-encoded Project room dataset.



(a) Image (b) Response Map

Figure 4.6 – An example showing (a) image and its corresponding (b) Harris response map.

over all the possible images pairs for each scene (Light Room and Project Room), i.e., each lighting condition of a scene is, in turn, the reference and the other conditions of that scene are the test images. Thereafter, we compute the average of RR over all such possible image pairs (52 pairs for Project Room and 42 pairs for Light Room) for a particular scene. Similar to [12, 82], our keypoint detection scheme is confined to the strongest 1000 keypoints. This is mainly to ensure a fair comparison of the blob or corner keypoint detection, as different detectors result in a highly different number of keypoints. We use a fixed detection error rate, i.e. $\epsilon = 3$px, which is .03% of image size (see section 5.4.5).

CC vs RR based optimization. We observe from Figure 4.4(a),(b) that CCGTM and CCBTM records substantial gap in performance with respect to RRGTM and RRBTM using both Harris and SURF detectors. We can conclude that high correlation between reflectance image pairs does not directly guarantee the stability of keypoint detection. However, this can be explained as follows. Keypoints are localized using non-max suppression technique as explained in [64]. The magnitude per pixel depicts the probability of that pixel to be detected as a keypoint which generally rely on second order derivatives. An example of response map using Harris detector [45] is shown in Figure 4.6, where high pixel values indicate higher likelihood to be considered as keypoint. As a consequence, local keypoints are in general sparingly distributed in detailed areas of an image. This process is much more complex than a simple statistical correlation computed at the pixel level. This is further illustrated in Figure 4.5. It can be observed that correlation of response maps are linearly proportional to RR whereas in contrast the correlation of the reflectance image maps shows scattered behavior. Therefore, this leads to the major conclusion of this study. In order to design a keypoint-detection-optimal tone mapping, the traditional Retinex based approaches need to take into account the detector response maps while estimating the reflectance images for a given scene.

Comparison with Traditional TMOs. From Figure 4.4(c),(d), we show that optimizing the traditional models in Eqs (4.1) and (4.2), RRGTM and RRBTM respectively, lead to large performance gains in terms of RR when compared to existing local and global TMOs. It shows the necessity to optimize the tone mapping operators with respect to detection tasks.

Finally, we also observe that the performance gains are significantly larger for Project Room than Light Room. This is mainly due to the fact that lighting transformations are much tamer for Light Room dataset, which entails smaller performance variations in RR when comparing different TMOs.

4.3 Learning a TMO for Efficient Keypoint Detection

The experimental study in the previous section concludes that optimizing TMO parameters with respect to RR leads to higher keypoint stability over the per-pixel similarity between the tone-mapped images. Though this study points to the parametric sensitivity in TMOs,

it does not provide any keypoint-detection-optimized TMO model. Therefore, the problem of designing an optimal TMO for the keypoint detection task remains open.

In this section, we therefore develop a novel learning-based adaptive tone mapping operator, referred as DetTMO, aims at enhancing keypoint stability under drastic illumination variations. To this end, we design a pixel-wise adaptive TMO which is modulated based on a model derived by Support Vector Regression (SVR) using local higher order characteristics. Our idea is mainly motivated by the conclusions of our previous study in Section 4.2 where optimizing tone mapping parameters for keypoint detection is shown to yield significant gains in RR. However, in that study optimal TMO parameters are computed globally on the whole image using grid search and, more importantly, for a given scene. Here, instead, we propose to *learn TMO parameters based on the local features* of the scene. Specifically, since keypoints are sparsely detected and depend on their neighborhood properties, we argue that local parametric modulations in TMOs can enhance the keypoint detection probability by adaptively mapping pixels based on their local higher-order characteristics.

To predict such optimal modulations in this context, we are inspired by the success of regression-based "task-optimization" models. In the literature, regression-based models have been explored for several image processing problems [53, 75, 102]. Here, we employ SVR, which has been successfully used, *e.g.* in image super-resolution [75], and which enables to cope with large variability in the input training samples compared with low-dimensionality approaches using explicit functions such as polynomial regression.

Additionally, we propose a simple detection-similarity-maximization model to generate appropriate training samples for the SVR. We initially consider several HDR image pairs which are taken with the same viewpoint with different lighting variations. Then, we define our objective function to find optimal modulation maps for such scenes so that the similarity of the detector response maps is maximized. For the defined objectives, the optimization is carried out using stochastic gradient descent (SGD) [10] by deriving the required partial derivative architecture.

In the following, we initially provide the details of our learning-based adaptive TMO approach, the similarity maximization model for generation of training set, the SVR training and proposed dataset. Later, we present the experimental results and analysis.

4.3.1 General Framework

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 a pixel x, TMO operates as: $I'(x) = \varphi(I(x), \boldsymbol{\theta})$, where $I(x) \in \Re$, $I'(x) \in [0, 255]$ and $\boldsymbol{\theta}$ represents a vector of parameters.

For several existing TMOs [21, 27, 67, 90], parameter $\boldsymbol{\theta}$ serves diverse objective such as filter shape and size, brightness control, but all motivated for visual perception. Such parameter is often set as globally for an image and further chosen by trial and test procedures. For example, $\boldsymbol{\theta}$ serves as variance in ChiuTMO [21], sharpening constant in

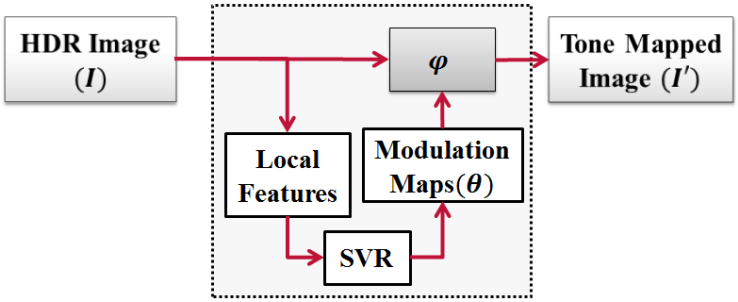


Figure 4.7 – *Learning based DetTMO.*

ReinhardTMO [90] and range and spatial variance in bilateral filtering based TMO [84].

Based on these observations, we assume function φ as an extension of existing tone mapping functions which can be modulated spatially by adapting their vector of parameters. The idea here is to facilitate the local adaption of function φ at sparse keypoint locations to further ease their identification and detection. In this chapter, we call the corresponding vector of parameters as modulation maps so as to distinguish their purpose of modulating the TMO locally from global parametric tuning. The modulation maps are given as $\boldsymbol{\theta}(x) = \{\theta_1(x), \theta_2(x)..,\}$ and our proposed TMO operates as: $I'(x) = \varphi(I(x), \boldsymbol{\theta}(x))$.

To predict the *modulations maps*, we propose to learn a model by employing SVR [99] while complying with the following two constraints: (a) To distinguish the keypoint and its neighborhood locations, (b) To bring invariance (as much as possible) to the non-affine lighting variations in the physical world scenes.

By using the radial basis kernel mapping, our SVR minimizes the non-linear problem of predicting modulation maps $\boldsymbol{\theta}$ by linearly separating the input samples in high-dimensional space. We refer the reader to [99] for more details about kernel-based SVR optimization model. Fig. 4.7 outlines the general framework of our proposed keypoint optimal TMO.

4.3.2 Adaptive Tone Mapping Operator

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 [21, 84]. Following this idea, our tone mapping function φ is expressed as: $\varphi = I \cdot L^{-1}$, where the illumination component L is estimated by an adaptive version of bilateral filtering [104] and is given as:

$$L(x, \boldsymbol{\theta}) = \frac{1}{W} \cdot \sum_{y \in \Omega} G_{\theta_1(x)}(\|x - y\|) \cdot G_{\theta_2(x)}(\|I(x) - I(y)\|) I(y), \tag{4.4}$$

where G is a Gaussian kernel. Here, modulation vector $\boldsymbol{\theta}$ 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 neighborhood set Ω and the normalization factor $W = \sum_{y \in \Omega} G_{\theta_1(x)}(\|x - y\|) \cdot G_{\theta_2(x)}(\|I(x) - I(y)\|)$.

It is important to note that we have built our model using the bilateral filtering, mainly because its proposed adaptive formulation facilitates the integration of local modulation in the proposed TMO. Moreover, it has been previously studied in the context of keypoint detection in HDR imaging in varying lighting conditions as discussed in Section 4.2. However, any other tone mapping techniques with parametric formulations such as [21, 90] could be used as well with our proposed framework.

4.3.3 Generation of Training Set: Detection Similarity Maximization Model

Suppose we are given a set of HDR scenes where each scene has images captured from the same viewpoint but with different lighting conditions. To train the SVR for our proposed model, we need to compute the "ideal" modulation maps (θ_1, θ_2 in our case) for a scene which ensures high keypoint stability. In other words, for a scene undergoing lighting variations, we need to estimate the modulations ensuring maximum keypoint repeatability. To this end, one solution is to design an optimization model which maximizes the RR of multiple images of a given sequence.

RR is a measure of detector efficiency, as defined in the previous section. Since RR is a non-smooth and non-differentiable function, it cannot be directly used to define the similarity objective of our optimization model. Therefore, we instead propose an alternative solution to use *differentiable* detector response maps $\mathcal{R}$ and design a model that maximizes the similarity between these response maps of image pairs drawn from a given sequence. $\mathcal{R}$ is a score map which determines a pixel's strength to be a keypoint and it mainly depends on the choice of keypoint detection algorithm.

Our response map $\mathcal{R}$ is generated by a Harris corner detector [45]. It is based on the autocorrelation scores computed per pixel using the second-order moment matrix, and is given as:

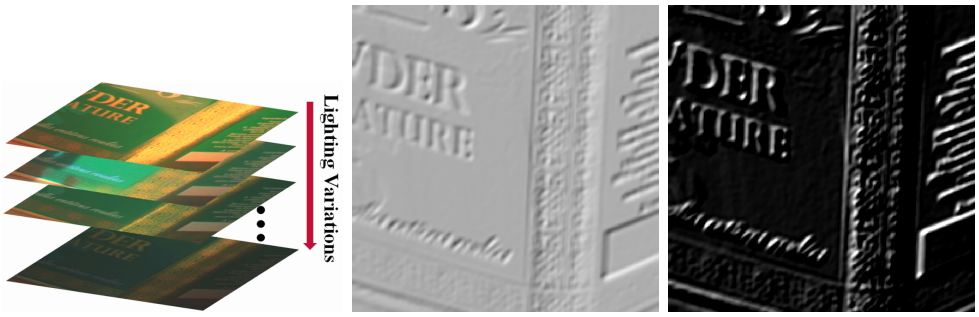
$$\mathcal{R}\big(\varphi(x,\boldsymbol{\theta})\big) = det\{\boldsymbol{M}\big(\varphi(x,\boldsymbol{\theta})\big)\} - k \cdot tr\{\boldsymbol{M}\big(\varphi(x,\boldsymbol{\theta})\big)\}^2 \tag{4.5}$$

where $\boldsymbol{M}$ is the second order moment matrix as detailed in [45]. k is the sensitivity factor ($k = 0.04$). Here, we have focused on the corner-based detectors as they are computationally inexpensive and highly used for real time applications, *e.g.* tracking, wide-view panorama creations, etc. However, the model could be extended to region or blob-based detectors as well.

Objective: Let S be a scene consisting of N HDR images with lighting variations as shown in Fig. 4.8 (a). Let $P = \big\{(1,2),(2,3)..,\big\}$ be the set of $K = \binom{N}{2}$ pair combinations of N images. Our aim is to maximize the response similarity by minimizing the following objective function:

$$\mathcal{F}(\theta) = \frac{1}{K} \sum_{\{i,j\} \in P} \Phi(\mathcal{R}_i(\theta), \mathcal{R}_j(\theta)), \tag{4.6}$$

and obtain the resulting modulation maps $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$ as shown in Fig. 4.8 (b) and (c).



(a) HDR Images (b) Modulation (θ_1) (c) Modulation (θ_2)

Figure 4.8 – *Generation of training set.* The samples images undergoing different lighting variations shown in (a) are used to generate the θ_1 and θ_2 modulation maps in (b) and (c) respectively, using the detection similarity maximization model.

Inspired by max-margin formulations for image retrieval tasks [87], we define function Φ using the logistic penalty

$$\Phi(\mathcal{R}_i, \mathcal{R}_j) = \log(1 + \exp(\epsilon - \langle \mathcal{R}_i \cdot \mathcal{R}_j \rangle). \tag{4.7}$$

where ϵ is a 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.

Optimization using SGD. We optimize the objective function in Eq. (4.6) using Stochastic Gradient Descent (SGD) [10]. To do so, we build the partial derivative architecture required for the SGD implementation as follows.

To estimate $\boldsymbol{\theta}$ maps at each iteration t, SGD update rule is given as:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_t \cdot \nabla \Phi_{\{i,j\}t}(\boldsymbol{\theta}_t), \tag{4.8}$$

where γ_t is a learning rate that can be made to decay with t as $\gamma_t = \gamma_0/(t+1)$ and the gradient for the objective function in Eq. (4.6) is replaced (as detailed in [10]) with the gradient of a randomly chosen sample pair $\{i, j\}$ at time t, *i.e.*

$$\nabla \Phi_{\{i,j\}}(\boldsymbol{\theta}_t) \triangleq \left. \frac{\partial \Phi(\mathcal{R}_i, \mathcal{R}_j)}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}_t}. \tag{4.9}$$

We computed the gradient required in Eq. (4.8) using the chain rule as follows,

$$\nabla \Phi_{\{i,j\}}(\boldsymbol{\theta}) = \left\{ \frac{\partial \Phi}{\partial \mathcal{R}_i} \cdot \frac{\partial \mathcal{R}}{\partial \varphi_i} \cdot \frac{\partial \varphi_i}{\partial \boldsymbol{\theta}}, \frac{\partial \Phi}{\partial \mathcal{R}_j} \cdot \frac{\partial \mathcal{R}}{\partial \varphi_j} \cdot \frac{\partial \varphi_j}{\partial \boldsymbol{\theta}} \right\} \tag{4.10}$$

4.3.4 Support Vector Regressor Training for DetTMO

An illustration for SVR training is shown in Fig 4.9. Let's assume that a scene with multiple images captured under different lighting variations is given for training. Further,

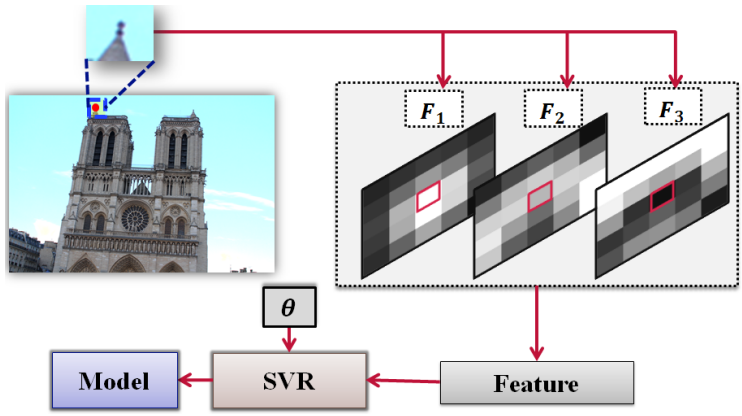


Figure 4.9 – *Training an SVR.* The sample pixel (red) with $s \times s$ neighborhood (blue) is chosen to extract the features maps (F_1, F_2, F_3) using response scores, gradients and intensity patterns respectively.

Figure 4.10 – Sample images from *HDR dataset*. The *HDR Dataset* is composed of 8 scene from different indoor/outdoor locations.

assume that the optimal modulation maps for the same scene are also given as described in Section 4.3.3.

To train an SVR model invariant to illumination variations, we first select random samples from keypoint and neighborhood locations across all the given images with varying lighting conditions. For each sample, we then consider a local patch of size $s \times s$ centered at that pixel. Next, we compute a feature vector which includes: a) the second-order detector response scores F_1, b) the gradient magnitudes F_2 and c) the local intensity patterns F_3.

The second-order response scores are based on the choice of the keypoint detector. Therefore, our response score feature for each pixel x in patch $s \times s$ is given as: $F_1(x) = det\{\boldsymbol{M}(x)\} - k \cdot tr\{\boldsymbol{M}(x)\}^2$. The gradient magnitudes for each pixel in the local patch is computed as: $F_2(x) = \sqrt{G_x^2(x) + G_y^2(x)}$, where G_x and G_y are the gradients in horizontal and vertical directions. The local intensity patterns for each patch is recorded by subtracting the value of centered pixel from other pixels and given as: $F_3(x) = I(x) - I(c)$, where c is the pixel at center location.

These individual features are normalized and concatenated to form the final feature vector $\{F_1, F_2, F_3\}$ of dimension $3s^2$ representing a training sample.

4.3.5 Luminance change HDR dataset

We propose an HDR dataset with 8 different HDR scenes as shown in Fig. 4.10. The *Light Room*, *Project Room* and *Poster* are the publicly available datasets and have been used for evaluating HDR for keypoint detection problems [11, 82]. However, these 3 scenes have been captured in indoor locations and hence, they are less challenging in terms of physical-world illumination transformations such as day/light change. Therefore, we captured 5 additional scenes including 1 indoor *Camroom* and 4 famous outdoor locations in Paris: *Notre-Dame*, *Louvre*, *Invalides* and *Grande Arche*. The *Camroom* scene is shot with a Canon Mark III camera in the presence of powerful 2K Watt reflectors. All the other outdoor HDR scenes are captured with Canon 700D camera at different times of the day. To create the HDR images, LDR images have been fused using the algorithm in [24]. Note that all scenes are geometrically calibrated.

4.3.6 Experimental Setup

We test our proposed model for keypoint detection task on 8 HDR scenes. We initially compare our DetTMO with the non-adaptive bilateral filtering based tone mappings BTMO and its globally optimized version BTMO(opt) as discussed in Section 4.2. Similar to our tone mapping function, both these TMOs are based on illumination normalization where the luminance L is estimated using bilateral filter. However, both these TMOs use global range and spatial variances. Moreover, BTMO(opt) is a variant of BTMO with an additional step of global parameter optimization, and approximates the maximum possible RR that can be achieved with BTMO model.

Then, we compare our model with state-of-the-art perception based TMOs: ChiuTMO [21], DragoTMO [27], ReinhardTMO [90] and MantiukTMO [67]. We considered these TMOs as they have been previously applied for HDR evaluation studies [101] for similar keypoint detection task.

We evaluate all these TMOs using popular and widely used corner detection schemes: Harris [45], Shi-Tom [104], FAST [94] BRISK [61]. In addition, even if our formulation is optimized for corner detection, we also test our TMO with respect to blob detectors such as SURF [8] and SIFT [64]. Since our model is designed for one image scale, we employed single-scale implementation for all keypoint detection schemes to ensure a fair comparison.

The detection performance is measured in terms of RR (as discussed in Section 4.3.3) with an error rate of 5 pixels. Namely, a keypoint is considered to be repeated in the test image if it lies in a circle of radius 5 centered on the projection of the reference keypoint onto the test image.

Training and Implementation details

For each test scene, we build the training set with 10,000 samples and use it to train and validate the SVR model. This training set is drawn from other scenes excluding the corresponding test scene. For instance, to test the Project Room scene, we build the training set by randomly selecting the samples from all other 7 scenes. For each training sample, we compute feature on a small patch size of 5×5 while following the feature extraction procedure from Section 4.3.4. Higher patch-size is not advisable as pixel correlation diminishes with increasing distance. Conversely smaller patch-size may extract insufficient information.

Implementation. We use the SVR implementation of LibSVM [15] using the Radial Basis Function (RBF) kernel. To obtain the optimal values of SVR parameters, the regularization cost and epsilon-SVR are tuned by 5-fold cross validation from the range of $[2^{-5}, 2^{15}]$ and $= [2^{-10}, 2^{5}]$, respectively.

We use the HDR Toolbox [7] for the implementation of the considered TMOs. Moreover, we use the Matlab's Computer Vision toolbox for Harris, Shi-Tom, FAST, BRISK and SURF, and Vlfeat for SIFT. Similar to previous keypoint evaluation studies [11, 82], we selected the strongest 500 keypoints from each test image.

4.3.7 Evaluation Results

Quantitative Results: We perform a thorough evaluation of our proposed DetTMO in quantitative terms as shown in Fig. 4.11 and Fig. 4.12. We basically evaluate the performance of our method over all test scenes using the Harris corner detector. In Fig. 4.11, we compare our model with the other variants of bilateral filtering based TMOs: BTMO and BTMO(opt) (in Section 4.2). These results clearly show that local modulation of bilateral filtering based

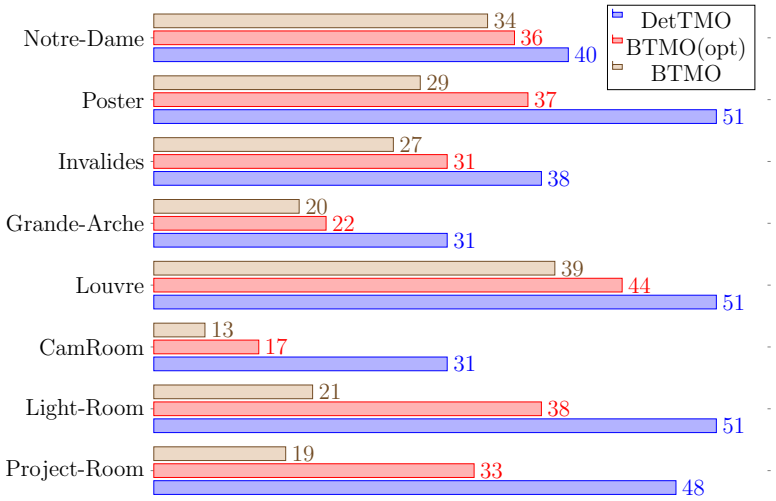


Figure 4.11 – *Quantitative Results I*: Repeatability Rates (RR) computed using DetTMO, BTMO(opt) and BTMO for each test scene using Harris keypoint detector. Note that while testing DetTMO for a particular scene we assured that the training for DetTMO is done on all other scenes.

tone mapping function using the proposed learned model significantly improves the keypoint stability across both the indoor and outdoor scenes.

Comparison with popular TMOs. We evaluate the performance of our method across different keypoint detection schemes including both corner and blobs. In Fig. 4.12, we initially compute the RR for all scenes for each considered TMO and then average them to compute the Average Repeatability Rate (AvgRR). We observe that for either detector (corner or blob) our proposed model outperforms all the other TMOs (perception based or keypoint-based). Further, the lower standard deviations observed with our proposed TMO shows higher stability of keypoints than other perception-based TMOs. Although our algorithm has been optimized for corners, it gives comparable or better performance with respect to other methods on blob detectors. This is partially due to the single scale implementation of the blob detectors used in this evaluation. However, the performance may differ when the multi-scale blob detection is taken into account.

We compare our DetTMO with popular and visually pleasing Reinhard TMO [90] and MantiukTMO [67]. In Fig. 4.13, we show that our method produces the highest number of repeated keypoints, even though both Reinhard TMO [90] and MantiukTMO [67] produce more visually appealing images.

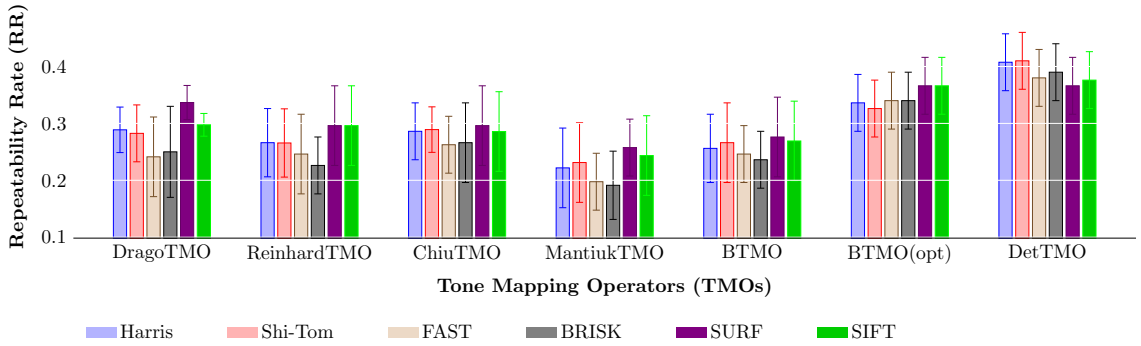


Figure 4.12 – *Quantitative Results II*: Average Repeatability Rates (AvgRR) computed on different TMOs using various keypoint detection schemes. The average is calculated over all test scenes.

Figure 4.13 – *Repeated Keypoints*. Row I: 2 HDR images from *Invalides* scene taken at different daytime. HDR images are displayed after log scaling[27]. Row II: the repeated keypoints using our proposed DetTMO (66 repeated keypoints out of strongest 200 keypoints). Row III: the repeated keypoints using Reinhard TMO (7 repeated keypoints out of strongest 200 keypoints). Row IV: the repeated keypoints using MantiukTMO (5 repeated keypoints out of strongest 200 keypoints).

4.4 Conclusions

In this chapter, we first investigate the impact of two factors for the optimization of tone mappings models. Build on its observations, later, we propose a new learning based adaptive tone mapping framework aiming at enhancing the keypoint detection performance under drastic lighting change scenarios. To that end, we train a Support Vector Regressor using local characteristic features to learn a model which spatially modulates the proposed pixel-wise adaptive TMO. Further, we introduce a simple and effective method for generating the training set to learn the SVR for the given problem. We evaluate our model on our proposed HDR benchmark dataset of indoor/outdoor scenes. Our model significantly outperforms state-of-the-art TMOs on the HDR dataset and also achieve state-of-the-art results across different keypoint detection algorithms.

Our current tone mapping model is designed for keypoint detection only. In the following, we plan to design a descriptor based tone mapping framework optimal for invariant descriptor extraction.

The work presented in this chapter has resulted in the following publications:

1. A. Rana and G. Valenzise and F. Dufaux, "Optimizing Tone Mapping Operators for Keypoint Detection under Illumination Changes", *2016 IEEE Workshop on Multimedia Signal Processing (MMSP 2016)*, Montréal, Canada, 2016.

2. A. Rana and G. Valenzise and F. Dufaux, "Learning-based Adaptive Tone Mapping for Keypoint Detection", *The International Conference on Visual Communications and Image Processing (ICME)*, Hong Kong, China, 2017.

Chapter 5

Learning a Tone Mapping Operator for Efficient Image Matching

5.1 Overview

Conventional TMOs have found to be sub-optimal for the feature extraction task, which includes a detection and a description stage. So far, we leveraged on learning the keypoint characteristics to design an optimal TMO for a stable detection only. In this chapter, we address the full feature extraction pipeline, including the description stage, to design an optimal TMO for efficient image matching.

More specifically, the goal of this chapter to find an optimal TMO which can enhance the extraction of stable features for scenes under complex real-world illumination transitions, such as day/night change. To this end, the chapter first proposes a descriptor-optimal TMO design, referred to as DesTMO, which solely aims at the extraction of invariant (as much as possible) descriptors from high-contrast areas of the scenes. Later, we introduce an optimal TMO, OpTMO, for full feature extraction chain (including both detectors and descriptors) which simultaneously enhances the detection rates and matching of features extracted from HDR scenes. Both the proposed task optimal TMOs namely, DesTMO and OpTMO, follow a learning based paradigm similar to the DetTMO in Chapter 4, but with entirely different design objectives.

Altogether, this chapter proposes

- a descriptor-optimal DesTMO which facilitates the extraction of luminance invariant descriptors.
- a locally adaptive, image-matching-optimal OpTMO which collectively address the detection and description stages of the feature extraction pipelines.

- an efficient method for generating appropriate training samples to circumvent the difficulty to train SVR in the context of DesTMO and OpTMO respectively. Additionally, we propose their differentiable surrogate objective functions.
- an evaluation of DesTMO and OpTMO against state-of-the-art methodologies. Furthermore, we show an applicative scenario of object localization.

5.2 Descriptor Optimal Tone Mapping Operator (DesTMO)

Our design idea is motivated by the detector optimal TMO of the previous chapter where significant gains in Repeatability Rate [70] were observed when optimal TMO parameters (controlling TMO's shape and size) were learned pixel-wise. However, we mainly focused on designing a tone mapping model for corner-like keypoint detection task, while here we consider a different problem, i.e., an optimal TMO for the extraction of discriminative descriptors.

To design DesTMO, initially a tone mapping function is introduced, which can be locally modulated by spatially varying its parameters. Its parameter maps are predicted by means of a learned illumination-invariant guidance model. Our guidance model is driven by the SVR and relies on the *gradient orientation-based features* that are extracted from densely sampled patches from the HDR content.

Unlike corner detection, descriptor extraction depends on the large set of neighborhood pixel-set (or patch) which are processed altogether to formulate the discriminative unique signature. Hence, we propose to *learn* the TMO parameters locally but based on *patch-level* information from the scenes. Specifically, since each descriptor is restricted to a patch size such as 16×16 in SIFT and SURF, we learn the TMO parameters on *patches* of the same size.

Since there is no standard dataset to train or test any model for DesTMO, we propose a simple 'descriptor similarity-maximization' approach to generate appropriate training samples. The objective function aims to maximize the similarities of descriptors if they are extracted from images from the same location but with lighting variations.

5.2.1 Proposed Model

Fig. 5.1 outlines the framework of our proposed algorithm. It primarily consists of a tone mapping function φ which maps the linear-valued HDR content of an image I to an output LDR I'. Similar to Chapter 4, it is expressed as

$$I'(x) = \varphi(I(x), \boldsymbol{\theta}), \tag{5.1}$$

where $I \in \Re^{m \times n}$, I' is of size $m \times n$ with pixel values in the $[0, 255]$ range, and $\boldsymbol{\theta}$ represents a vector of modulation maps, $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$, where θ_k is of size $m \times n$. Secondly, the framework

consists of a guidance model where an SVR predicts the optimal values of these modulation maps $\boldsymbol{\theta}$ by using the densely extracted local features from the HDR content. To this end, initially, the HDR image is densely sampled into patches of size $s \times s$ and from each such patch a SIFT feature f is extracted. Then, these features are fed to the regressor which in turn predicts parameter values for modulation map θ_1, θ_2. Note that the regressor output for each feature is applied over the size $s \times s$ in these modulation maps, corresponding to exact location of the sampled patch from which the feature is extracted. Such patch level tuned vector parameters θ_1, θ_2 are later used by φ to obtain the tone mapped image I'.

5.2.2 Tone Mapping Function

Inspired by illumination normalization TMOs [21, 29, 84] and similar to Chapter 4, our tone mapping function φ in Eq. (5.1) is expressed as: $\varphi = I \cdot L^{-1}$, where the illumination component L is estimated by a variant of bilateral filtering [104] and is given as:

$$L(x, \boldsymbol{\theta}) = \frac{1}{W} \cdot \sum_{y \in \Omega} \mathcal{G}_{\theta_1(x)}(\|x - y\|) \cdot \mathcal{G}_{\theta_2(x)}(\|I(x) - I(y)\|) I(y), \tag{5.2}$$

where $\mathcal{G}$ is a Gaussian kernel and for each pixel location x, the pixel y is in the neighborhood set Ω. The normalization factor W is equal to $\sum_{y \in \Omega} \mathcal{G}_{\theta_1(x)}(\|x - y\|) \cdot \mathcal{G}_{\theta_2(x)}(\|I(x) - I(y)\|)$. Here, the modulation vector $\boldsymbol{\theta}$ has two components: θ_1 and θ_2. They are often globally referred to as *spatial* and *range* variance respectively and control the behavior of function φ. For example, if θ_2 is predicted higher at a patch location, its corresponding Gaussian kernel widens and flattens behaving like a Gaussian blur [104], and finally, a blurred luminance L is estimated. In such condition, the final tone mapped pixels, which are obtained by normalizing the estimated L for the corresponding patch, will preserve the structures such as gradients.

Notice that we opted for bilateral filtering because its proposed formulation facilitates the integration of the core concept of local modulation. However, any other tone mapping function with parametric formulations such as [21, 90] could be used.

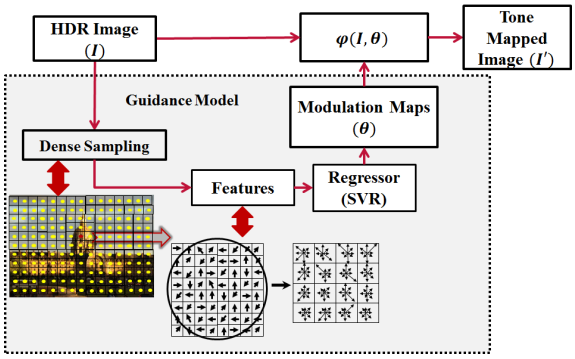


Figure 5.1 – *DesTMO*. The architecture of our proposed TMO.

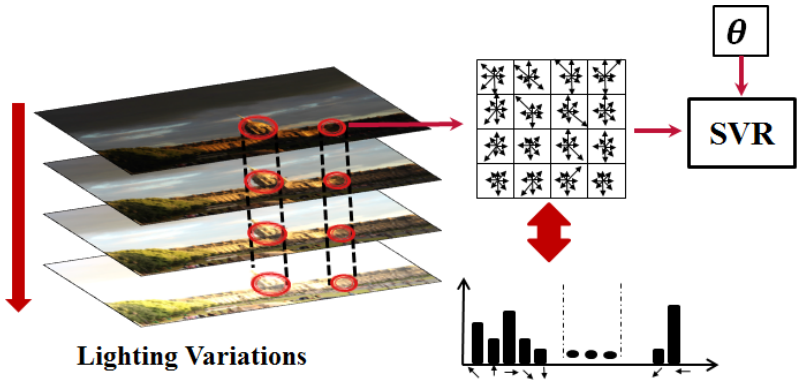


Figure 5.2 – *Training Pipeline*

5.2.3 Guidance Model based on SVR for DesTMO

We consider a training set $\{(f_1, o_1), ..(f_n, o_n)\}$, where f_i is the feature sample and o_i represents its corresponding observation (scalar or vector), $i = 1...n$. A classical linear regressor would solve the problem of fitting a prediction function as: $r(f_i) = (\omega^T f_i + b)$, where ω, b are estimated by minimizing the mean square error. However, such function is often incapable of separating the non-linearly sampled data, like our case where f_i is the SIFT feature with size 128, and $o(i) = \theta_{k(i)}$, where $k = 1, 2$. Therefore, with such given inputs, we use the non-linear SVR [99] which maps the input vector f_i into high dimensional space using the kernel ψ where data becomes linearly separable and is given as $r(f_i) = (\omega^T \psi(f_i) + b)$. To fit the desired non-linear SVR prediction function, the following optimization problem is solved:

$$\min_{\omega,b,\xi,\xi^*} \frac{1}{2}\|\omega^2\| + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

subject to:

$$\theta_{k(i)} - (\omega^T \psi(f_i) + b) \leq \chi + \xi_i,$$
$$(\omega^T \psi(f_i) + b) - \theta_{k(i)} \leq \chi + \xi_i^*,$$
$$\xi_i, \xi_i^* \geq 0, i = 1..n$$

where ξ, ξ^* are the slack variables, C represents the cost which is imposed for samples that exceed the error χ. For further understanding of the non-linear SVR optimization problem, we refer the reader to [99].

5.2.4 Generation of Samples

To train the SVR, we need to find appropriate training features and their corresponding *supervised* observations θ_1, θ_2 as shown in Fig. 8.9. To this end, we propose a two step solution. First, we identify *key* locations in a scene, where we can extract meaningful descriptor features. Second, we build a model to find the optimal θ_1 and θ_2 that maximize

the similarity between those descriptors which are captured from the same key locations of the scene.

To identify key locations, we first detect keypoints independently in each log-scaled HDR image of the scene using the DoG [64] detector. We then iteratively check, for each detected keypoint, whether it is found at about the same location in other images of the same scene, taken under different illumination conditions. If it is detected in the majority of these images, we call it a *key* location. As we just want to collect 'meaningful' *key* locations with majority occurrence under lighting variations, any other format could also be used instead of log-HDR.

From each key location, we use SIFT [64] as training feature, extracted from linear HDR content. More specifically, it is given as concatenation of 16 unnormalized cells *i.e.*$[\mathbf{x}_1, ...\mathbf{x}_{16}]$ where each cell can be compactly defined as [25, 109]:

$$h(\Theta|p)[\mathbf{x}] = \int \mathcal{G}_\delta(\Theta - \angle\nabla p(y))\mathcal{G}_{\hat{\sigma}}(y-x)\|\nabla p(y)\|d(y) \tag{5.3}$$

where $\mathbf{x}$ is the center location of the cell in the restricted square patch p of size 16×16. The independent variable Θ represents the gradient orientation ranging from $[0, 2\pi]$. Moreover, $\mathcal{G}$ represents the Gaussian kernel with standard deviation $\hat{\sigma}$ and an angular dispersion parameter δ.

Similarity model: We assume a scene S consisting of n HDR images with lighting variations as shown in Fig. 8.9. We consider $P = \big\{(1,2),(2,3)..,\big\}$ to be the set of $K = \binom{N}{2}$ pair combinations of N descriptors extracted from a key location. Our aim is to minimize the following objective function:

$$\mathcal{F}(\boldsymbol{\theta}) = \frac{1}{K}\sum_{\{i,j\}\in P} \Phi(\boldsymbol{h}_i(\boldsymbol{\theta}), \boldsymbol{h}_j(\boldsymbol{\theta})). \tag{5.4}$$

We define function Φ using the logistic penalty (similar to max-margin formulations in [87]),

$$\Phi(\boldsymbol{h}_i, \boldsymbol{h}_j) = \log(1 + \exp(\epsilon - \boldsymbol{h}_i^T \boldsymbol{h}_j). \tag{5.5}$$

We optimize the objective function in Eq. (5.4) using a robust optimization technique, Stochastic Gradient Descent (SGD) [10]. SGD update rule to estimate θ maps at each iteration t is given as: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_t \cdot \nabla\Phi_{\{i,j\}t}(\boldsymbol{\theta}_t)$, where γ_t is a learning rate which decays with t as $\gamma_t = \gamma_0/(t+1)$ and the gradient for the objective in Eq. (5.6) is replaced (as detailed in [10]) with the gradient of a randomly chosen sample pair $\{i, j\}$ at time t, *i.e.*$\nabla\Phi_{\{i,j\}}(\boldsymbol{\theta}_t) \triangleq \frac{\partial\Phi(\boldsymbol{h}_i,\boldsymbol{h}_j)}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_t}$.

Figure 5.3 – Scenes from *HDR luminance dataset*. The dataset is composed of 8 scene from different indoor/outdoor locations.

5.3 Results and Discussion

5.3.1 Experimental Setup

We build the test setup for image matching using the HDR luminance dataset shown in Fig. 5.3 which consists of 4 indoor and 4 outdoor scenes as detailed in Chapter 4. We compare our proposed DesTMO with the classical perception based TMOs: BTMO [84], ChiuTMO [21], DragoTMO [27], ReinhardTMO [90] and MantiukTMO [67].

The BTMO in [84] and ChiuTMO [21] are also based on normalizing the estimated luminance L but use global parametric settings. DragoTMO [27] maps the HDR content based on adaptive logarithmic scaling. ReinhardTMO [90] and MantiukTMO [67] are well known tone mapping techniques for high visual quality outputs with appealing brightness and contrast. We considered these TMOs as they have been previously applied for HDR evaluation studies [84, 101] for the related task of feature detection.

To effectively evaluate the impact of descriptor extraction scheme, we selected the strongest 500 keypoints using the DoG detector [64] for each tone mapped image. Then, we use four popular and widely used SURF [8] and SIFT [64] descriptor schemes as well as the FREAK[77] and BRISK [61] binary descriptors.

Metrics: We evaluate the descriptor performance using the standard measures of Matching Score and mAP rates as detailed in Section 2.2.3. To define a match, we use the standard nearest neighbor distance ratio (NNDR) matching strategy.

Training and Implementation details

For each test scene, we build the training set with 5000 training samples and use it to train and validate the SVR model. Given a test scene from our dataset Fig. 5.3, the training set is drawn from the other 7 scenes. For each training sample, we compute the SIFT feature on a patch size of 16×16.

Implementation. We use the SVR implementation of LibSVM [15] using the Radial Basis Function (RBF) kernel. To obtain the optimal values of SVR parameters, the regularization cost and epsilon-SVR are tuned by 10-fold cross validation from the range of $[2^{-5}, 2^{15}]$ and $[2^{-10}, 2^{5}]$, respectively. We use the HDR Toolbox [7] for the implementation of the considered TMOs, Matlab's Computer Vision toolbox for SURF, FREAK, BRISK and

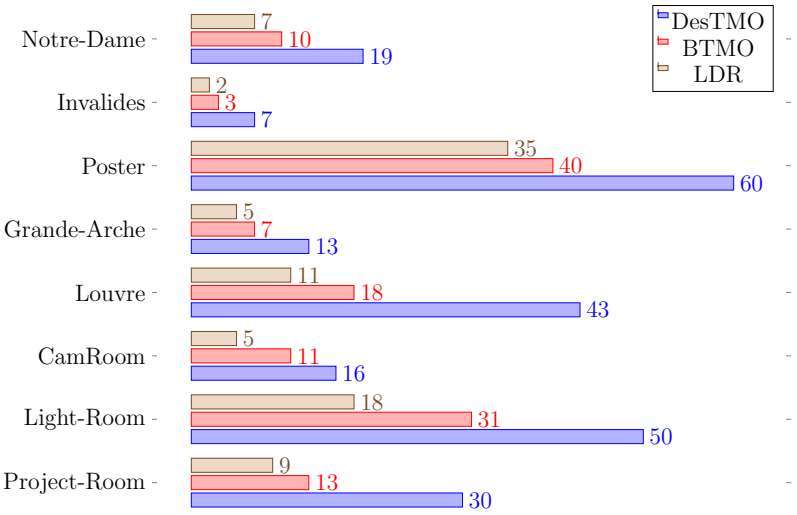


Figure 5.4 – **Matching Score** computed using DesTMO, BTMO and LDR for each test scene using SURF descriptor.

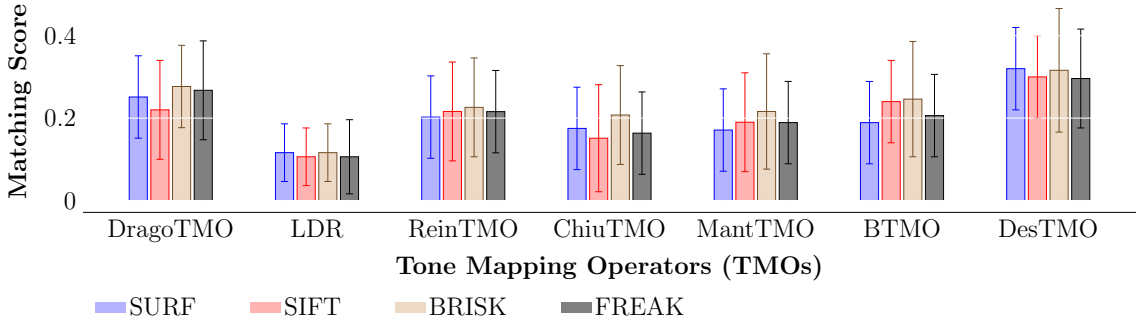


Figure 5.5 – **Average Matching Scores** computed on different TMOs using SURF, SIFT, FREAK, BRISK descriptor extraction schemes. The average is calculated over all test scenes.

Vlfeat [109] for SIFT.

5.3.2 Evaluation Results

We perform a thorough evaluation of our proposed DesTMO quantitatively using the matching score and mAP. We initially show in Fig. 5.4 the performance of our method over all test scenes using the SURF descriptor, where we compare our algorithm with BTMO [84] and the best exposure LDR. Our results clearly show that predicted local modulation of the bilateral filtering helps in preserving the invariance of the local gradient and hence, boosts the average number of correct matches in both the indoor and outdoor scenes. However, we observe small gains in outdoor scenes such as Invalides. This can be explained by strong lighting transitions and is partially due to increased false matches due to repetitive structures in the images as shown in Fig. 5.7. Note that, we use threshold $th = 0.2$ to avoid ambiguous matches and to improve the readability of descriptor matching

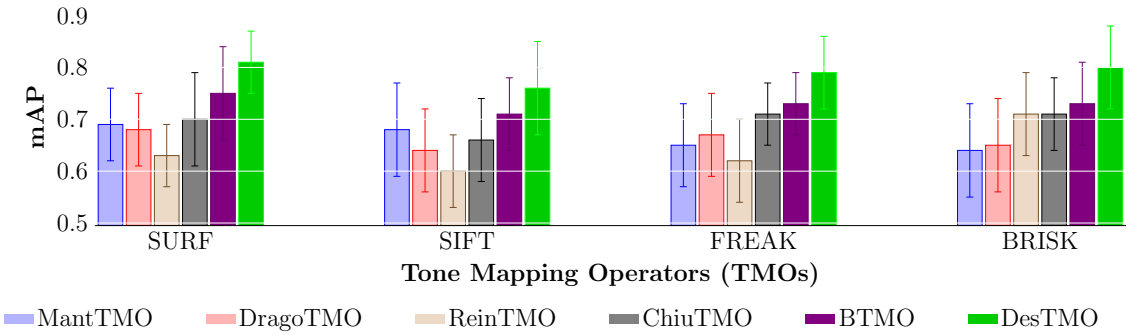


Figure 5.6 – **Mean Average Precision** (mAP) rates computed on different TMOs using SURF, SIFT, FREAK, BRISK descriptor extraction schemes. The average is calculated over all test scenes.

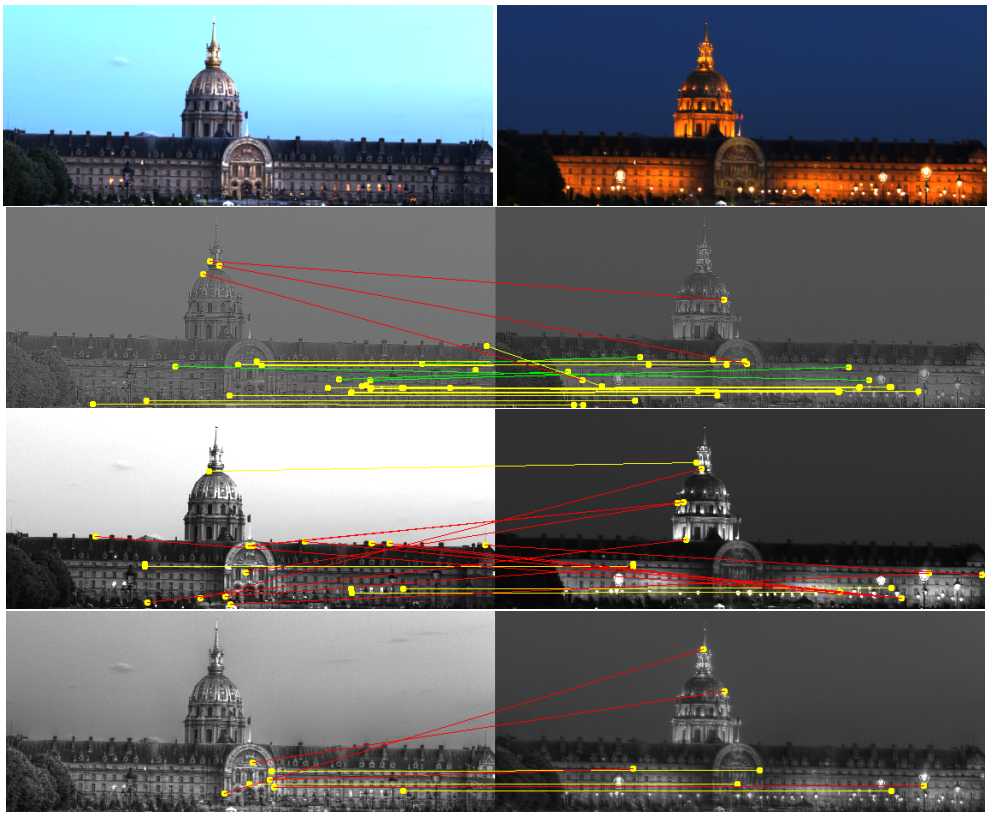


Figure 5.7 – Day/Night matching using SURF. Row I: 2 HDR images from *Invalides* scene are displayed after log scaling[27]. Correct and incorrect matches are shown with yellow and red lines respectively. Green lines represent the special case of mismatch due to repetitive structure. Row II: the feature matching using our proposed DesTMO (11 correct and 3 incorrect matches). Row III: using Reinhard TMO (3 correct and 11 incorrect matches). Row IV: using MantiukTMO (4 incorrect and 3 correct matches).

in Fig. 5.7.

Comparison with popular TMOs. We evaluate the performance of our method across different descriptor extraction schemes including both gradient based and binary descriptors. In terms of average matching score, we observe that for every extraction scheme,

our DesTMO yields a higher number of correct matches, as shown in Fig. 8.6. Furthermore, in Fig. 5.6, we compute the mAP rates by averaging the area-under-the-curve of PR curves of the complete dataset. We observe that for every descriptor extraction scheme, our proposed model outperforms all the other TMOs. Additionally, we compare our proposed TMO with popular and visually pleasing Reinhard TMO [90] and MantiukTMO [67] in Fig. 5.7, where we show that our method produces a higher number of correct matches in difficult day/night matching.

5.4 Optimal Tone Mapping Operator for Image Matching

In the previous Chapter, we designed a detector-optimal DetTMO, controlled by a guidance model which is *learned* to understand the keypoint's locally extremal and covariant characteristics. Similarly, we introduced a descriptor-optimal DesTMO where the guidance model is mainly trained to facilitate the invariant densely-sampled descriptor extraction. However, both these TMOs only handle one aspect at a time, namely, keypoint detection or descriptor extraction. This is inefficient in practice for the image matching task, e.g., a poor detector degrades descriptor matching [70].

Notice that optimizing a TMO considering keypoint detection and description concurrently is not trivial, as the corresponding design objectives are generally different and somehow contrasting. For instance, an optimal TMO for detection aims to produce covariant feature points, while a TMO optimal for description should guarantee some form of invariance to transformations over a local neighborhood. In addition, optimal detection requires an accurate localization of keypoint position, while optimal description is a patch-level process. In Chapter 3 [83], we have showed that TMOs that are optimal for detection are not necessarily so when the full matching chain is considered.

In this section, we introduce 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. For OpTMO, we initially introduce a tone mapping function similar to DetTMO, which can be locally modulated by varying spatially (pixel-wise) its parameters as a function of the HDR content characteristics. Afterwards, we propose a *guidance model* to map HDR-based local characteristics features (detection and description-based) to a low-dimensional TMO parameter space, by means of a support vector regressor (SVR) [99].

For OpTMO, 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 the proxy cost functions which enables to compute the required derivatives and obtain an optimal solution. In the following, we detail each step.

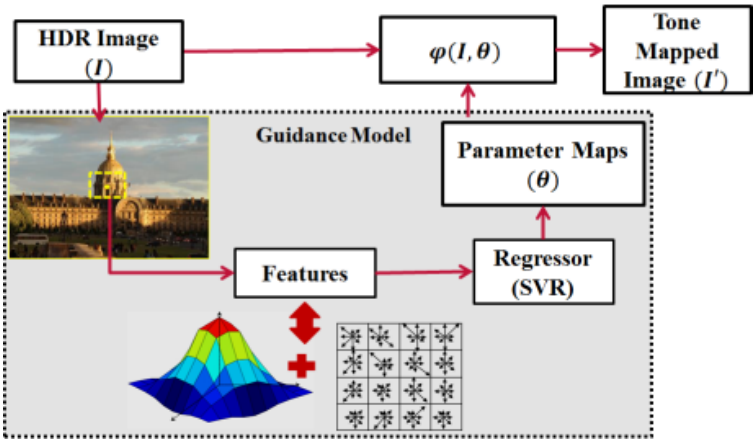


Figure 5.8 – Optimal Tone Mapping Design. The tone mapping function is modulated by the SVR-based guidance model, which predicts optimal parameter maps using the characteristic features.

5.4.1 Optimal Tone Mapping for Image Matching

Fig. 8.4 outlines the general framework of our proposed optimal TMO for image matching. It primarily consists of a tone mapping function φ which maps the linear-valued HDR content of an input image I to an output LDR I'. Our choice of the tone mapping function is same as described in 5.2.2. Secondly, the framework consists of a guidance model where a learned SVR predicts an optimal parameter map $\boldsymbol{\theta}$ based on local HDR content characteristics. In the following subsections, we discuss the design of our proposed framework and how to generate training data for the SVR.

5.4.2 Generation of Training Set

In this section, we address the problem of generating an adequate ground truth set for training the SVR-based model. We aim to find such ground truth parameter maps $\boldsymbol{\theta}$, which result in efficient image matching (*i.e.* mAP score) for a scene which undergoes drastic lighting variations, as shown in Figure 5.9. In this section, we, therefore, formulate an objective function f, which we minimize over the $\boldsymbol{\theta}$ to yield the optimal parameter maps. The proposed total energy f represents the difference in the image matching pairs. We quantify this difference in terms of both keypoint detection and descriptor extraction stages, depicted as 'Detection Response' and 'Description', respectively, in Figure 5.9. Finally, we propose to optimize the objective using the SGD based optimization method to obtain the optimal $\boldsymbol{\theta}$.

In the following, we first discuss the formulation of the objective function f. Then, we detail the considerations with respect to image matching components in view of designing the objective function f. Finally, we detail the SGD-based method to optimize the objective to obtain the optimal $\boldsymbol{\theta}$.

Objective Function

We aim to optimize $\boldsymbol{\theta}$ to tone map an image for the full feature extraction pipeline. Therefore, the objective function should consolidate each stage of the feature extraction pipeline *i.e.* to locate and extract the feature. Henceforth, we introduce two energy terms dedicated to keypoint localization (E_{det}) and descriptor extraction (E_{des}), respectively and define a combined objective function as:

$$\underset{\theta}{\text{minimize}} \quad f(\theta) = E_{det}(\theta) + E_{des}(\theta). \tag{5.6}$$

where each energy term is computed over a scene consisting of N HDR images with lighting variations as shown in Fig. 5.9 (a). Let $P = \big\{(1,2),(2,3),\dots\big\}$ is the set of $K = \binom{N}{2}$ pair combinations of N images. The E_{det} term aims to ensure the covariance 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. Both terms are detailed as follows:

The Energy term E_{det}

To ensure efficient matching, we observe that it is important to enforce the similarity in detection response maps [84]. This is mainly because high similarity response maps increases the probability of detection of keypoints at similar locations and thereby enhances the probability of correct matches.

We define the detection similarity term E_{det}, by summing the penalty computed from each pair in the set K, as:

$$E_{det} = \frac{\lambda_{det}}{K} \sum_{\{i,j\} \in P} \mathcal{C}_1(\mathcal{R}_i(\theta), \mathcal{R}_j(\theta)). \tag{5.7}$$

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

$$\mathcal{C}_1(i,j) = \log(1 + \exp(\epsilon_c - \langle \mathcal{R}_i \cdot \mathcal{R}_j \rangle), \tag{5.8}$$

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 which belongs to a scene, and $\langle \cdot \rangle$ denotes the scalar product. The selection of $\mathcal{R}$ is detailed later in this section.

Inspired by the max-margin formulations applied to retrieval [87] or classification tasks [98], we use the logistic function as the penalty in our detection term. It is a smooth differential operator and ideally penalizes less if there is high similarity and vice-versa. Note that the term E_{det} is somewhat similar to the one we proposed in detector optimal TMO in Chapter 4. But, in this work, we include an additional factor λ_{det} which weights

the penalization corresponding to detection.

Selection of Response : From handcrafted [96] to deep-learning [116] era, the concept of corner-like keypoint detection methods has gained popularity for low-latency vision tasks due to high speed, less computational complexity and competitive accuracy. By definition, corners exhibit low correlation with neighboring pixels in all directions. The most basic and widely adopted corner detectors [36, 45, 104] localize the extrema primarily by computing the per pixel gradient autocorrelation matrix, given as:

$$\boldsymbol{M} = \begin{bmatrix} I_x^2 & I_{xy} \\ I_{yx} & I_y^2 \end{bmatrix}, \tag{5.9}$$

where each component represent the directional derivative. Thereafter, different methods are proposed in the literature to localize the extrema "keypoints" [96]. In this work, we use [45] which describes the response for each pixel x without directly computing the eigenvectors of $\boldsymbol{M}$ as:

$$\mathcal{R}(\boldsymbol{x}) = det\{\boldsymbol{M}(x)\} - k \cdot tr\{\boldsymbol{M}(x)\}^2, \tag{5.10}$$

where k is tuned empirically.

Similar to the baseline as discussed in Section 4.3.3, we employ the detector response in Eq. (5.10) mainly because it is based on the popular structural matrix $\boldsymbol{M}$, which is simpler to differentiate than alternative approaches, thus aiding in backpropagation. Note that alternate detection methods could also be used, but our choice has been made entirely based on the computation complexity and ease of use in backpropagation.

The Energy term E_{des}

The energy term E_{des} aims to penalize the dissimilarity of the descriptors extracted from the tone-mapped images. Previously in DesTMO, we proposed a densely sampled patch-based method where a model is learned to predict global parametric values for an individual patch. Hence, not only the method optimized $\boldsymbol{\theta}$ for a patch *globally*, but it also lacked the consideration of keypoint localization. In contrast, the image matching pipeline additionally relies on the localization of the descriptors. Hence, here we argue that it is important to compute the gradient orientation impact per pixel and to focus on its locations prior to designing a descriptor-based penalty function. It not only helps in preserving the salient locations but also avoids any "look-alike" redundant matches [86]. Therefore, we propose to constraint the penalization to the dissimilarity of those descriptors that belong to some potential *keypoint* region. We define E_{des} as:

$$E_{des} = \frac{\lambda_{des}}{K} \sum_{\{i,j\} \in P} \mathcal{C}_2(\mathcal{D}_i(\theta) - \mathcal{D}_j(\theta)), \tag{5.11}$$

where $\mathcal{C}_2$ is the Euclidean distance and λ_{des} is a weighting factor. To apply the constraint in practice, we compute the descriptor $\mathcal{D}$ after the keypoint localization which is obtained by applying the softargmax operation $\mathcal{S}$ [16] on the resulting response map. In general terms, $\mathcal{S}$ is given as

$$\mathcal{S} = \sum_i \frac{\exp(\beta z_i)}{\sum_j \exp(\beta z_j)} \cdot i \tag{5.12}$$

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. Cluttering refers to a phenomenon when several keypoints are located close to each other [86].

To compute an accurate keypoint localization, we define the final gradient orientation around each pixel location computation as follows:

$$\mathcal{D} = \begin{cases} h(\nu|p), & \text{if } \mathcal{S}(\mathcal{R}) \geq \Lambda \\ 0, & \text{otherwise} \end{cases} \tag{5.13}$$

where $h(\nu|p)$ is the gradient orientation feature map explained later in Eq. (5.14) 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. (5.11).

Selection of h : A common image matching approach relies on the similarity of features extracted from patches corresponding to detected keypoint locations. One widely used descriptor extraction algorithm is the Scale Invariant Feature Transform (SIFT) [64] which is a concatenation of 16 unnormalized cells *i.e.* $[\mathbf{c}_1, ...\mathbf{c}_{16}]$, where each cell can be compactly defined as [25, 109]:

$$h(\nu|p)[\mathbf{c}] = \int \mathcal{G}_\delta(\nu - \angle\nabla p(y))\mathcal{G}_{\hat{\sigma}}(y - c)\|\nabla p(y)\| d(y), \tag{5.14}$$

where $\mathbf{c}$ is the center location of the cell in the restricted square patch p of size 16×16. The independent variable ν represents the gradient orientation ranging in $[0, 2\pi]$. Moreover, $\mathcal{G}$ represents the Gaussian kernel with standard deviation $\hat{\sigma}$ and an angular dispersion parameter δ. Once histograms are computed, they are normalized and concatenated into a single 128-dimensional descriptor.

SGD Implementation Details

We optimize the objective function in Eq. (5.6) using stochastic gradient descent (SGD) [10]. It is a fast and robust optimization technique to estimate the incremental gradient descent by its stochastic approximation using a randomly chosen sample from the initial set. To

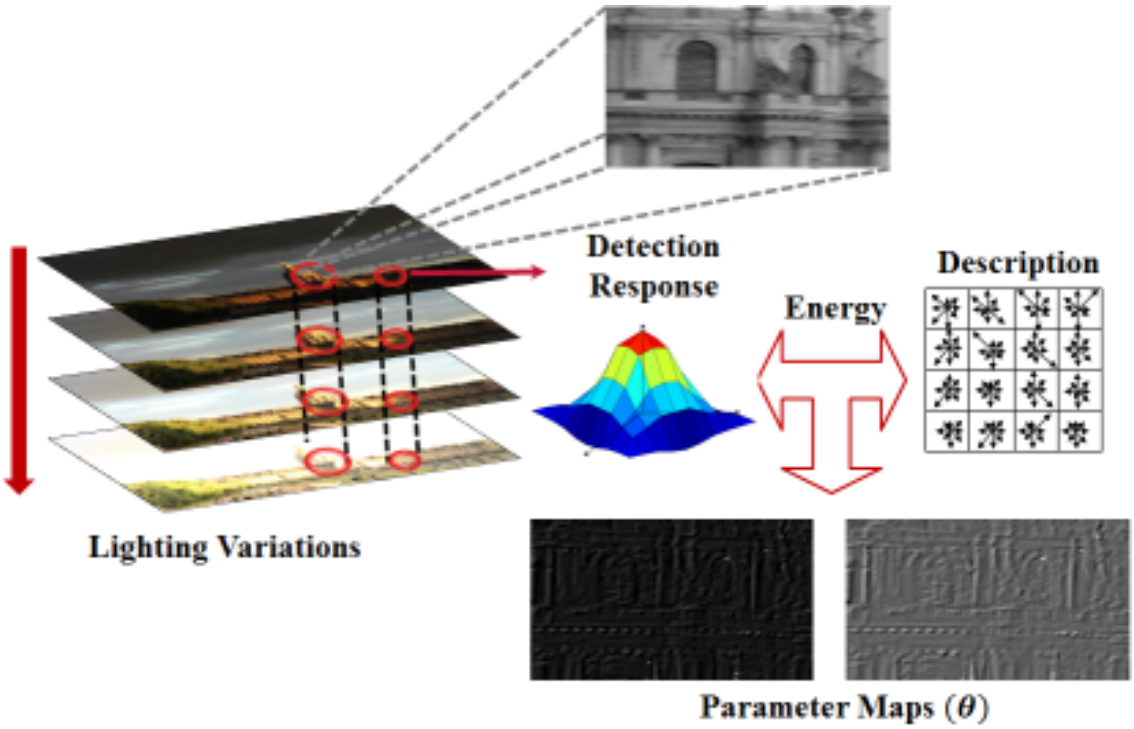


Figure 5.9 – Generation of Training Set. Ground-truth parameter maps are generated by minimizing the total energy determined from a set of images of the same scene, undergoing lighting variations, using the procedure in Section 5.4.2.

implement the SGD based optimization, we follow the backpropagation procedure. We initially build the required partial derivative framework with the objective given in Eq. (5.6). It is more formally expressed as

$$\nabla \boldsymbol{\mathcal{C}}_{\{i,j\}}(\boldsymbol{\theta}) = \left\{ \frac{\partial \mathcal{C}_1}{\partial \mathcal{R}_l} \cdot \frac{\partial \mathcal{R}_l}{\partial \varphi_l} \cdot \frac{\partial \varphi_l}{\partial \boldsymbol{\theta}} + \frac{\partial \mathcal{C}_2}{\partial \mathcal{R}_l} \cdot \frac{\partial \mathcal{R}_l}{\partial \varphi_l} \cdot \frac{\partial \varphi_l}{\partial \boldsymbol{\theta}} \right\} \Big|_{l=i,j} \tag{5.15}$$

Then, following the SGD rule, we iteratively estimate $\boldsymbol{\theta}$ by randomly selecting sample (i, j) from the set P. Finally, we compute the gradient of the objective in Eq. (5.6), that is:

$$\frac{\partial f}{\partial \theta}\Big|_{\theta_t} = \frac{1}{K} \sum_{\{i,j\} \in P} \frac{\partial \boldsymbol{\mathcal{C}}}{\partial \theta}\Big|_{\theta_t} \tag{5.16}$$

where

$$\boldsymbol{\mathcal{C}} = \lambda_{det} \cdot \mathcal{C}_1(\mathcal{R}_i, \mathcal{R}_j) + \lambda_{des} \cdot \mathcal{C}_2(\mathcal{D}_i, \mathcal{D}_j), \tag{5.17}$$

with the single (i, j) selected image pair. Thereafter, at each iteration t, SGD update rule is given as:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \boldsymbol{\gamma}_t \cdot \nabla \boldsymbol{\mathcal{C}}_{\{i,j\}t}(\boldsymbol{\theta}_t), \tag{5.18}$$

where γ_t is a learning rate is made to decay with t as $\gamma_t = \gamma_0/(t+1)$, and the gradient for the objective function in Eq. (5.6) is replaced by the gradient of a randomly chosen sample

pair $\{i, j\}$ at time t, *i.e.*

$$\nabla \boldsymbol{\mathcal{C}}_{\{i_t,j_t\}}(\boldsymbol{\theta}_t) \triangleq \left. \frac{\partial \boldsymbol{\mathcal{C}}(\mathcal{R}_{i_t}, \mathcal{R}_{j_t}, \mathcal{D}_{i_t}, \mathcal{D}_{j_t})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}_t}. \tag{5.19}$$

For SGD based optimization, we start from a randomly initialized set of $\boldsymbol{\theta}$ which are updated iteratively using the update rule in Eq. (5.18). In total, the model comprises 3 hyper-parameters: $\gamma_0, \lambda_{det}, \lambda_{des}$. To estimate these hyperparameters, we follow the standard approach used in [110] and take a small set of pairs from P and perform a simple cross-validation using the grid search method in the log scale. For the SGD related optimization and convergence proofs along with the aysmptotic analysis, we refer the reader to [10].

This proposed mechanism for finding the optimal parameters $\boldsymbol{\theta}$ for a function φ using SGD is generic, *i.e.* one can easily tune the parameter maps of any TMOs that can be expressed as Eq.(5.1).

5.4.3 Support Vector Regressor Training for OpTMO

Consider the sample set of characteristic features $\mathcal{F} = \{\boldsymbol{f_1}, \ldots, \boldsymbol{f_n}\}$ 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 $\boldsymbol{f_i}$ by concatenating two parts: a) the gradient-based SIFT pattern [64] (64 dimensional feature); and b) the 5×5 grid-based detector response feature [85] (25 dimensional feature). This forms a total dimension of 89. The features $\boldsymbol{f_k}$ 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 [12]. Finally, for each training sample, we get the following input-output corresponding pairs $\{(\boldsymbol{f_1}, \theta_{k(1)}), ..., (\boldsymbol{f_n}, \theta_{k(n)})\}$ and formulate our prediction problem using SVR. To fit the desired nonlinear SVR prediction function, the corresponding optimization problem is solved using the dual maximization approach.

5.4.4 Experimental Results and Discussion

Dataset : We consider the HDR dataset presented in our previous designs of DetTMO and DesTMO, which is composed of 8 different HDR scenes as shown in Fig. 5.3.

5.4.5 Evaluation Metrics

We evaluate the keypoint detection and descriptor extraction performance on the tone mapped images using the standard measures of Repeatability Rate (RR) and Matching Score (MS) respectively, as detailed in Section 2.2.3. For the evaluation of the full image matching, we compute the mean average precision (mAP) scores [70].

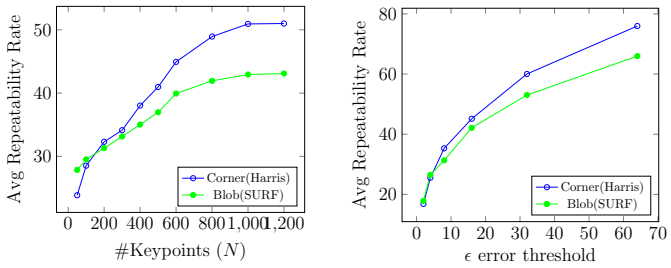


Figure 5.10 – Repeatability Rates (RR) computed for OpTMO using a corner (Harris) and a blob (SURF) keypoint detector.

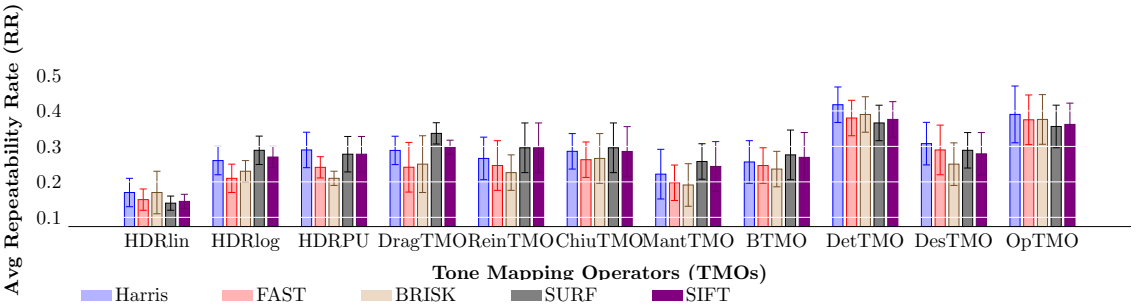


Figure 5.11 – **Keypoint Detection I**: Average Repeatability Rates (AvgRR) computed on different TMOs using various keypoint detection schemes. The average is calculated over all test scenes.

5.4.6 Evaluation Setup

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. For detection and description stage, we formulated a total of 280 test image pairs respectively from the 8 scenes.

We compare the proposed OpTMO with classical perception-based TMOs, including: BTMO [29], ChiuTMO [21], DragoTMO [27], ReinhardTMO [90] and MantiukTMO [67]. In addition, we also consider our previously proposed DetTMO [85] and DesTMO [86], which are optimized methods for detection and description only, respectively.

SVR Training and Implementation We use the SVR implementation of LibSVM [15] using the Radial Basis Function (RBF) kernel. The optimal values of SVR parameters, the regularization cost and epsilon, are obtained by 10-fold cross validation from the range of $[2^{-5}, 2^{15}]$ and $[2^{-10}, 2^{5}]$, respectively.

To train and validate the SVR model, we build the training set with 5000 sample feature set for each test scene. This training set is drawn from other scenes excluding the corresponding test scene. For instance, to test the Project Room scene, we build the training set by randomly selecting samples from all other 7 scenes. For each training sample, we randomly select a pixel location and compute characteristic features around the selected

location, while following the feature extraction procedure described in Section 5.4.3. A window of 16×16 is selected around the pixel location for computing the gradient based SIFT pattern part of the feature $\boldsymbol{f_i}$. Whereas a 5×5 grid based detector response is used for the second part of the $\boldsymbol{f_i}$.

5.4.7 Keypoint Detection

We evaluate all the considered TMOs using Harris [45], FAST [94], BRISK [61], SURF [8] and SIFT [64] (as detailed in Section 2.2). We selected these detection methods based on the state-of-the-art studies in evaluating the performance of TMOs [11, 19, 82, 86] and also due to their popularity in real time applications [107].

RR [70] is sensitive to the number of detected keypoints and the error rate ϵ. For instance, large variations in the number of keypoints across different scenes might lead to biased average scores. Therefore, we fix the keypoint detection to the strongest N keypoints as suggested in prior TMO evaluation studies [12, 82, 83]. The impact of N and ϵ over average RR score is shown in Fig. 5.10. Overall increase in number of keypoints leads to an increase in average RR but the growth slows down after a certain number, partially due to the detection of cluttered keypoints. On the other hand, increase in the average RR with the increasing ϵ is in coherence with the findings of [96]. Here, we choose the values $N = 500$ and $\epsilon = 10$.

Implementation We use the HDR Toolbox [7] for the implementation of the considered TMOs. Moreover, we use the Matlab's Computer Vision toolbox for Harris, FAST, BRISK and SURF, and Vlfeat for SIFT.

Comparison We perform a thorough evaluation of our proposed OpTMO quantitatively using the RR measure. In Fig. 8.12, we initially show the performance of our OpTMO and other state-of-the-art TMOs in terms of RR averaged over all test scenes. 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 [82, 83].

In Fig. 5.12, we expand our experimental test bench for each scene and compare the performance of our OpTMO with the globally optimized BTMO [84] and our previously proposed detector-optimal DetTMO [85]. The per scene gains of OpTMO over BTMO prove that local modulation of parameters significantly improves the keypoint stability. In addition, we observe that the gain in performance between local and global optimization depends significantly on content characteristics. Especially for indoor scenes, which have been acquired by varying locally the illumination and introducing stark shadows, local

parameter tuning enables to obtain important RR gains. We also notice that OpTMO achieves similar (within 2-4% per scene) RR as DetTMO, which is optimized for keypoint detection *only* and thus provides an upper bound in the achievable repeatability.

In order to further confirm these observations, we report a head-to-head comparison of OpTMO versus BTMO and DetTMO, respectively, in Fig. 5.13, for two different detectors: Harris (corner) and SURF (blob). OpTMO has higher RR whenever a point (representing a specific scene and illumination condition) is above the 45° line. As expected, we observe that this is often the case for BTMO, while for DetTMO the two methods have very similar performance. As mentioned above, the loss in keypoint repeatability compared to DetTMO is expected, and is mainly due to two reasons. On one hand, the additional descriptor-level cost term in Eq. (5.11) changes the objective function with respect to detector repeatability only (as in DetTMO). On the other hand, the use of the softargmax localization measure in Eq. (5.12) reduces cluttering of keypoints in our OpTMO. This is illustrated on a detail of the "Project-Room" scene in Fig. 5.14. For instance, cluttered keypoints are detected near the beaver's eyes in DetTMO, whereas OpTMO handles such detections efficiently. Interestingly, the composite objective function in Eq. (5.6) enables to achieve RR almost as good as DetTMO, but with a significantly improved descriptor matching and thus overall image matching performance, as shown in the next section.

Finally, we observe from Fig. 5.13 that these conclusions are valid for both Harris and SURF detectors, in spite of the fact that OpTMO is trained with respect to a classical corner response function (Eq. (5.10)). This demonstrates experimentally that images tone mapped with the proposed approach lead to increased detection performance even when the actual used detector is different from the specific response characteristics captured by the proxy cost function used for training. This is mainly because our models preserves the local neighborhood surrounding the 'extrema' which stabilizes the localization of keypoints by other detectors as well.

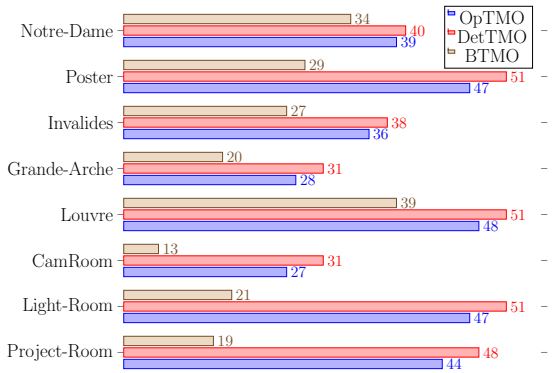
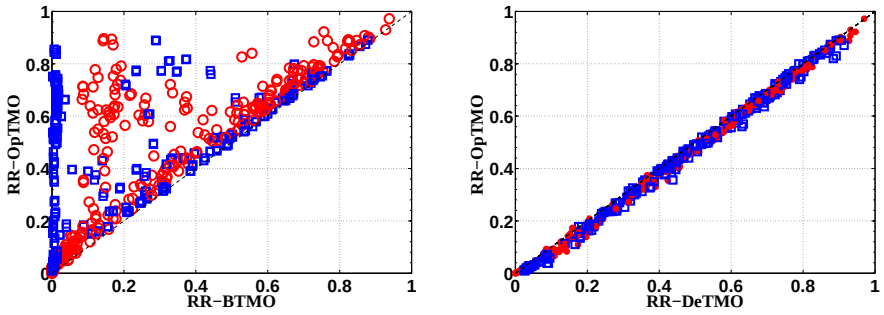
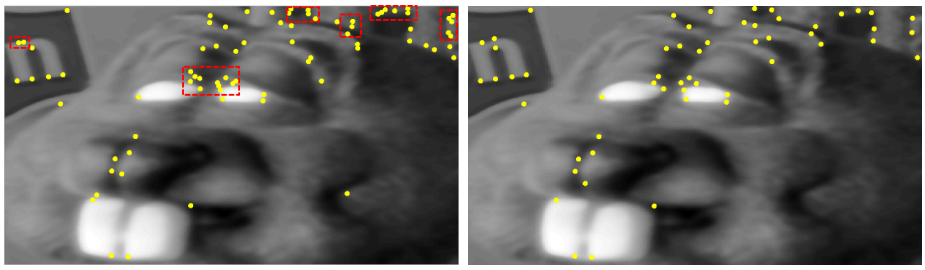


Figure 5.12 – **Keypoint Detection II**: Average Repeatability Rates (RR) computed using BTMO [84], DetTMO [85] and the proposed OpTMO for each test scene using Harris keypoint detector.



OpTMO vs BTMO OpTMO vs DetTMO

Figure 5.13 – **Keypoint Detection III**. The head to head comparison between (a) OpTMO vs BTMO and, (b) OpTMO vs DetTMO. Each point represent an image pair with different lighting conditions from the HDR dataset. The points represented using o depict the Harris corner detector and □ represents the SURF blob detector.



DetTMO OpTMO

Figure 5.14 – **Keypoint Detection IV**. Harris corner keypoints on the DetTMO and proposed OpTMO. The cluttered keypoints in DetTMO are highlighted using the red squares.

5.4.8 Descriptor Matching

We perform a thorough evaluation of our proposed OpTMO for descriptor matching using BRISK [61], FREAK [77], SIFT [64] and SURF [8] descriptors. We use the matching score (MS) as performance measure considering the NNDR matching criteria with a threshold value $th = 0.5$.

Implementation We use the Matlab's Computer Vision toolbox for FREAK, BRISK and SURF, and Vlfeat for SIFT, with their default parameter settings.

Comparison In Fig. 5.15, we compare the average OpTMO MS with respect to state-of-the-art TMOs. Overall, we attain significant gains in terms of MS using all feature extraction methods. With $th = 0.5$ (default value [64, 70]), the gains are considerable for gradient-based features schemes such as SIFT and SURF, which is expected by design given the definition of the descriptor signature in Eq. (5.13).

To further analyze these results quantitatively, in Fig. 5.16 we report per scene comparison between the competing TMOs that are observed from Fig. 5.15. We observe that

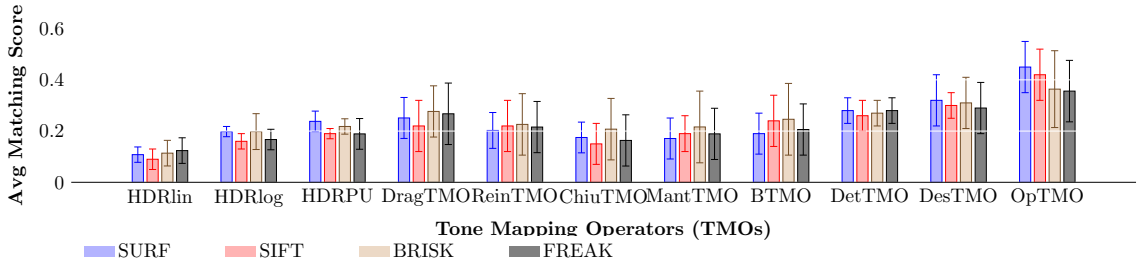


Figure 5.15 – **Descriptor Matching I** computed on different TMOs using SURF, SIFT, FREAK, BRISK descriptor extraction schemes. The average is calculated over all test scenes.

for each scene (indoor or outdoor) our OpTMO outperforms all the other TMOs. As in Section 5.4.7, we observe considerable gains with respect to traditional BTMO, confirming the potential of local parameter optimization. In comparison to DetTMO, we observe that gains are not as high as what are obtained with BTMO. This can be explained by the higher RR of the DetTMO (Fig. 5.12) which improves the probability of the correct matches. Interestingly, we also observe that in many scenes DetTMO and DesTMO perform equally well, e.g., *Invalides* and *Project-Room* scenes. This is mainly because DesTMO is not optimal for detection, which entails a higher number of false matches.

Finally, we show the per image-pair analysis in Fig. 5.17 to further analyze the behavior of individual test pairs. We observe that our OpTMO improves the MS over DesTMO across the whole dataset (i.e., the gains are not concentrated on specific image pairs). In fact, there is not a single case where there is a significant drop in OpTMO's performance against the descriptor-optimal DesTMO, which again confirms the advantages of simultaneously optimizing the TMO for keypoint detection and description. In addition, the OpTMO produces consistent gains even if a binary descriptor such as FREAK is employed, in spite of the use of a gradient-based cost function in Eq. (5.13).

Note that MS is sensitive to the choice of th. Therefore, in the following section, we perform a global image matching evaluation using mAP to overcome the impact of the threshold.

5.4.9 Image Matching

We evaluate the full image matching chain by computing mean average precision (mAP) scores over the complete dataset. We obtain the mAP rates by averaging the area-under-the-curve of PR curves [70]. The results per TMO are reported in Fig. 5.18. 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 Fig. 8.14, 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. From Fig. 8.14, we observe that optimizing only for detector response (DetTMO) might produce a higher number of false matches. On the other hand, optimizing with respect to descriptor matching only (DesTMO) cannot ensure high matching efficiency due to the lower keypoint repeatability. Instead, efficient image matching can only be ensured by optimizing the TMO with respect to the full feature extraction chain, as in the proposed OpTMO.

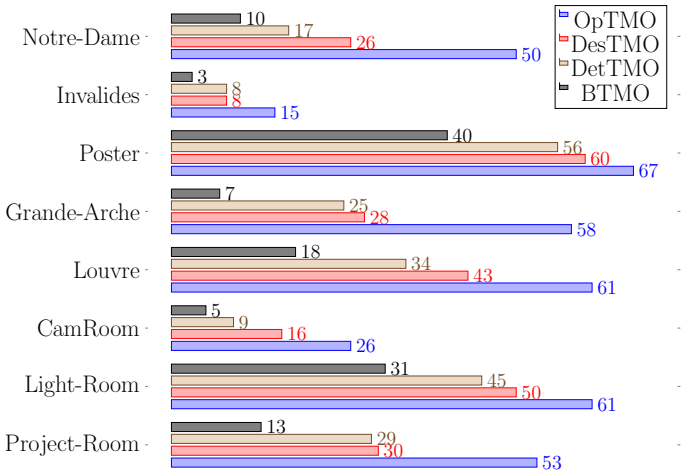


Figure 5.16 – **Descriptor Matching II**. Matching Score comparison between BTMO [84], OpTMO, DetTMO [85] and DesTMO [86] over all the scenes in the HDR dataset using SURF feature extraction scheme.

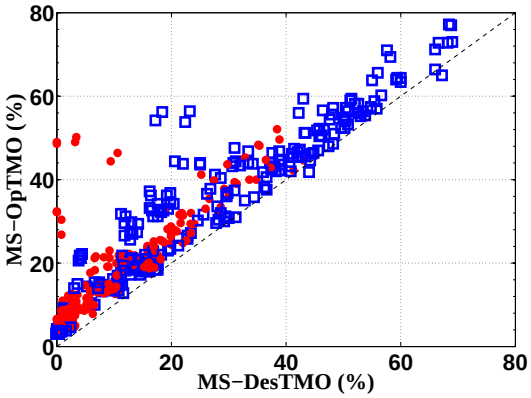


Figure 5.17 – **Descriptor Matching III**. Matching Score comparison between DesTMO vs OpTMO over all the scenes in HDR dataset. The points represented using o corresponds to FREAK feature detection scheme and □ corresponds to SURF scheme.

5.4.10 Applications

Localization of objects is a high-precision and pivotal task in many computer vision applications, *e.g.*to find region of interest for fine-grained recognition challenges. For localization, first a homography matrix is computed by finding the best matching correspondences

between the target and the test image. Then, the desired object is localized based on the estimated geometric relationship. In Fig. 5.20 and Fig. 5.21, we show a similar applicative scenario of localization of selected objects such as structures in images undergoing both lighting and rotational transformations. We compare the performance of our proposed image-matching optimal TMO and the widely used ReinhardTMO over three scenes, namely *Louvre*, *ProjectRoom* and *Notre Dame*. In Fig. 5.20, we first find the corresponding matches between the two scenes using the SURF scheme for each TMO. Then, based on those resulting matches, we estimate the homography as proposed in [46]. We observe that our model gives more correct corresponding matches in all three scenes as compared to ReinhardTMO. In challenging outdoor scene such as *Louvre* where there is a direct impact of sunshine, we observe that ReinhardTMO results in all incorrect matches, mainly concentrated in the brightest regions. In Fig. 5.21, we overlay the results on the test tone-mapped images to show where exactly our desired object should be located based on the obtained correspondences. In *Louvre* and *ProjectRoom* scenes, we observe that tone-mapped images using our proposed model result in correct localization of the desired object in the test image, as compared to ReinhardTMO. In the *Notre Dame* scene, the impact of illumination on the target region is smaller, and we are able to find correct overlaying results using both tone mappings.

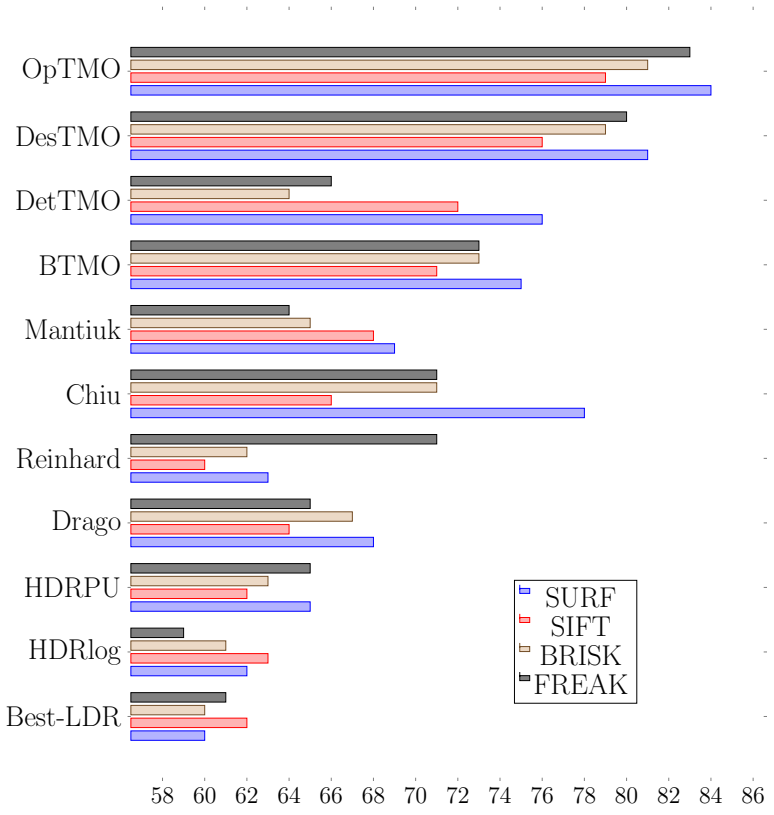


Figure 5.18 – **Image Matching I**. mAP % scores for the 9 different LDR modalities using 4 feature extraction schemes. Scores are averaged over 8 lighting change datasets.

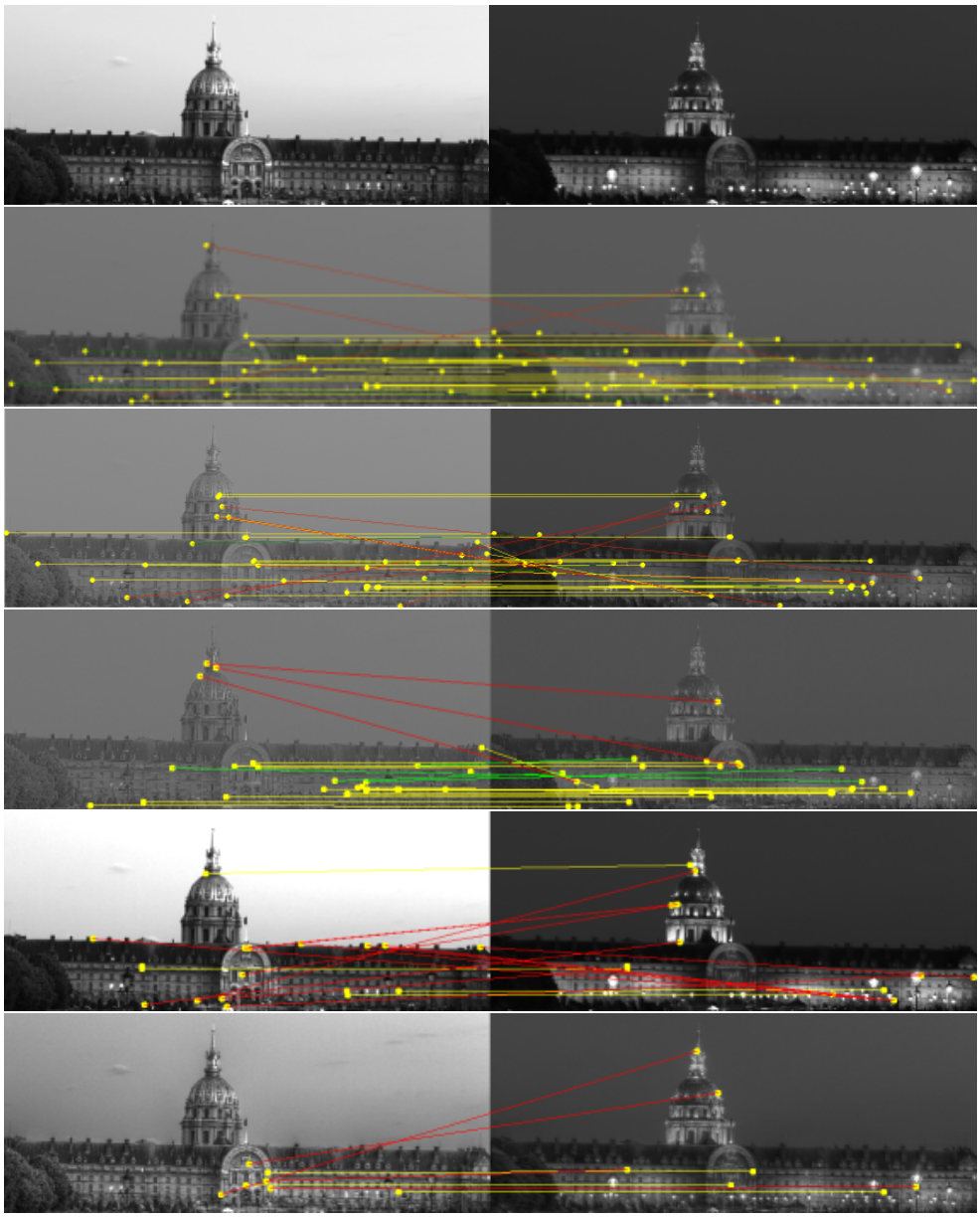


Figure 5.19 – **Image Matching II**. Day/Night matching using SURF. Row I: 2 HDR images from *Invalides* scene are displayed after log scaling [27]. Correct and incorrect matches are shown with yellow and red lines, respectively. Green lines represent the special case of mismatch due to repetitive structure. Row II: the feature matching using our proposed OpTMO (21 correct and 3 incorrect matches). Row III: using DetTMO (13 correct and 6 incorrect matches). Row IV: using DesTMOusing (11 correct and 3 incorrect matches). Row V using Reinhard TMO (3 correct and 11 incorrect matches). Row VI: using MantiukTMO (3 correct and 4 incorrect matches).

Computation Time: In Fig 5.22, we compare the execution time (i.e. to tone map an HDR image) of the most competing state-of-the-art TMOs namely, BTMO, DetTMO, DesTMO and OpTMO. The computational time of our proposed method is not very far from the DesTMO. Note that the current implementation has been carried on a Intel Xeon

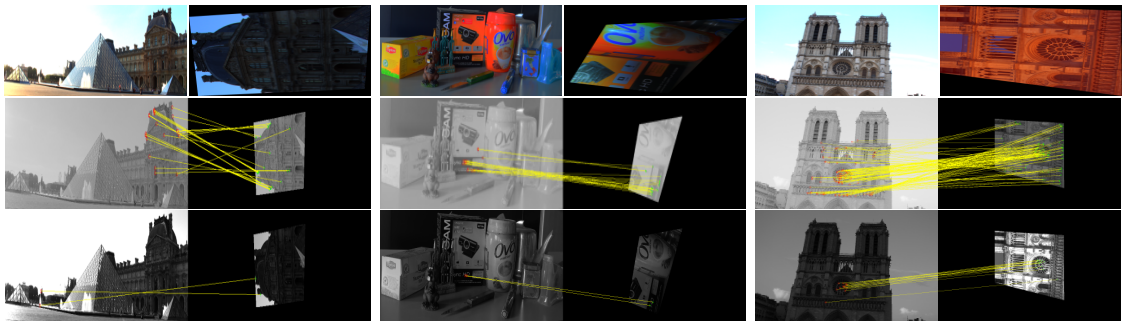


Figure 5.20 – **(Match & Locate)** Row I: Pair of HDR images from *Louvre*, *ProjectRoom* and *Notredame* scenes, with one reference and other being a selected region undergone lighting change and rotation. Row II: the feature matching using our proposed OpTMO. Row III: using Reinhard TMO.

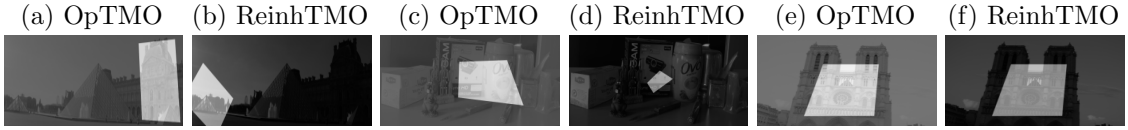


Figure 5.21 – **(Match & Locate)** Final patch localization results shown by overlaying the matched area for each scene using OpTMO and Reinhard TMO

CPU 4 cores processor, 16 Gb RAM windows 7 machine and has not been parallelized. An efficient parallelized implementation can further sped up the execution.

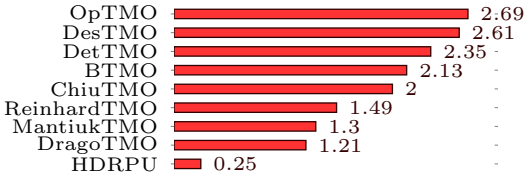


Figure 5.22 – *Computation time in sec (log scale)*. The time is computed by running all TMOs for an image size (512×512) on a Intel Xeon CPU 4 cores processor, 16 Gb RAM windows 7 machine.

5.5 Conclusions

We propose a novel task optimal TMOs to improve the descriptor discriminability and image matching efficiency under drastic changes of lighting conditions. To this end, we first generated training samples by proposing a objective function capturing both the detection and description stages of the feature extraction pipeline. Later, we trained a Support Vector Regressor using local characteristics to learn a model which predicts spatially varying TMO parameters. We evaluate the proposed OpTMO on a HDR dataset of indoor/outdoor scenes where it outperforms state-of-the-art TMOs across different image matching algorithms. Finally, we demonstrated the performance of our method over other TMOs in a simple localization based application scenario. Our proposed task-optimal TMO can be applied to

different detection/description approaches and can be directly fused with any local feature based applications such as structure from stereo, scene reconstruction, object tracking, recognition and photogrammetric applications.

So far, we build our optimal TMO model by proposing the variants of existing tone mapping functions and learning their parameter maps. Although we obtained gains on image matching task over the state-of-the-art, the proposed models are bounded by the limited functionalities of these exiting tone mapping functions. Hence, in the following, instead of learning the tone mapping parameters, we will focus on directly learning a tone mapping function.

The work presented in this chapter has resulted in the following publications:

1. A. Rana and G. Valenzise and F. Dufaux, "Learning-based Tone Mapping Operator for Image Matching", *IEEE International Conference on Image Processing (ICIP'2017)*, Beijing, China 2017.

2. A. Rana and G. Valenzise and F. Dufaux, "Learning-based tone mapping operator for efficient image matching", *IEEE Transaction of Multimedia(TMM)*, 2017 submitted, currently under revision.

Chapter 6

Deep Tone Mapping Opeartor for HDR Imaging

6.1 Overview

With a given task-based objective, so far we proposed models by relying on specific characteristics of a given tone mapping function. In fact, a variant of Bilateral filtering has been adopted to showcase model's learning ability. However, not all the TMOs are differentiable and consequently, difficult to be learned using the proposed methods. Moreover, an individual TMO addresses only some specific characteristics which might be desired based on the content. This raises a natural question whether a more general tone mapping function can be formulated which can be easily trained for any given task and adapt itself for all the real world scenes.

In this chapter, we address this question by designing a generic end-to-end TMO which adapts itself for all the real world scenes considering the desired task-specific characteristics. Leveraging on large HDR dataset for *perceptual* objectives, we propose the first deep learning based tone mapping (DeepTMO) architectural designs for converting a *linear* HDR content into a high resolution tone mapped LDR output. The current implementation of our models are trained for a perceptual task *i.e.* to give the most realistic and high quality output without any visible damage to its content. Since a large amount of labeled HDR image dataset is absent for designing a task-optimal deep-learning based TMO, the proposed architecture can additionally serve as a baseline for HDR based analysis. In future, this could be explored by fine tuning the proposed model with an in-line cascaded task specific deep learning model *e.g.* for image matching, face detection, video surveillance etc.

This chapter presents 3 distinctive deep learning based tone mapping networks namely DeepTMO-R, DeepTMO-S and DeepTMO-HD. Based upon conditional generative adversarial networks (cGAN) [43, 72], each of the proposed model directly takes in linear HDR content and reproduces a realistic looking image aiming to mimic the original HDR content with pixel values in the range [0-255]. Unlike conventional convolution neural

networks (CNNs) explored in previous HDR related works [30, 32, 41], our architecture avoids the requirement of explicitly defining a task specific loss function. This happens mainly because our networks are trained to model by themselves, loss functions adapted from the underlying training data.

To train these proposed models, we accumulate data from available HDR image sources. However, a major challenge while training the models arise from the absence of any publicly available training dataset. Selecting ground truth through a subjective evaluation over a large dataset is a highly tedious tasks. Thus, it necessitates the requirement of an objective quality assessment metric which can quantify the tone mapping performance of each TMO for any possible scene. For our task, we select a well known metrics namely Tone Mapped Image Quality Index (TMQI), which is used to rank 13 widely used tone mapping operators. Using this, for each HDR input, we select the one which ranks topmost on this objective metric score as our ground truth tone mapped output.

Finally, our DeepTMO implicitly learns the best characteristic of all available global, local and perceptually based TMOs over a wide variety of scenes. In a sense, both through its architectural design and the underlying dataset, it is conditioned to preserve global features (such as overall structures, contrasts and luminance) as well as local finer details (such as local texture patterns) thus yielding high quality visually pleasing tone mapped outputs.

In a nutshell,

1. We propose the first deep learning based tone mapping operator, which can generate visually pleasing realistic tone mapped outputs for a wide variety of HDR inputs.

2. We fully explore and compare 3 different cGAN architectures designed specifically for generating high resolution LDR tone-mapping outputs preserving overall structural information as well as local fine-grained details.

3. We overcome the challenge of unavailability of ground truth tone mapped images for our HDR dataset by utilizing an objective metric to quantify and rank various TMOs.

4. We provide an extensive comparison of our proposed methodology with thirteen different tone mapped operators over large dataset of 105 images.

5. Our proposed model can be explored for future deep learning based task-optimal TMO designs.

6.2 Deep Learning for HDR Image Analysis

Recently, deep Convolutional neural networks (CNNs) have been utilized extensively for multiple high dynamic range imaging tasks such as reconstructing HDR using a single

exposure LDR [30], predicting and merging various high and low exposure images for HDR reconstruction [32] or yielding HDR outputs from dynamic LDR inputs [52]. CNNs have also been modeled to learn an input-output mapping as done for de-mosaicking and de-noising by [40] or learning an efficient bilateral grid for image enhancement [17]. [41] have recently proposed a deep bilateral tone mapper, but it works only for 16-bit linear images and not the conventional 32-bit HDR images. A major drawback of all these past techniques is that, although the learning process is completely automated, one still needs to explicitly specify an effective loss function that the CNN learns to minimize. Thus the quality of resulting output is dependent a lot on the choice of our loss function. Formulating a loss function that constrains the CNN to yield sharp tone-mapped LDR from their corresponding linear-valued HDR is complex and an ill posed problem. Although authors in [30] formulated the inverse tone mapping task by formulating the problem in log domain, but in our case, we rather work in linear domain to avoid any form of information loss.

6.2.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [43] have attended lot of attention owing to their capability to model the underlying target distribution by forcing the predicted outputs to be as indistinguishable from the target images as possible. Through this, they implicitly learns an appropriate loss function, eliminating the requirement of hand crafting one by an expert. This property has enabled them to be utilized for a wide variety of image processing tasks such as super-resolution [60], photo-realistic style-transfer [51] as well as semantic image in-painting [115].

For our task, we employ GANs under a conditional setting, or better called as conditional GANs (cGANs) [72], where the generated output is conditioned on the input image. One distinctive feature of cGAN framework is that they learn a structured loss where each output pixel is conditionally dependent on one or more neighboring pixels in the output image. Thus, this effectively constrains the network by penalizing any possible structure that differs between output and target. This property is quite useful for the task of tone-mapping where we only want to compress the dynamic range of an HDR image, keeping the structure of output similar to our desired target. For this specific reason, cGANs have been quite popular for the task of image-to-image translation, where one representation of a scene is automatically converted into another, given enough training pairs [48] or in an unsupervised setting [63, 106, 120]. However, a major limitation of using cGANs is that it is quite hard to generate high resolution images due to training instability and other optimization issues. The generated images are either blurry or contain noisy artifacts. In [18], motivated from perceptual loss [51], a direct regression loss is derived to generate high resolution 2048×1024 images, but this method fails on preserving fine-details and textures. In [111] significant improvement have recently been shown on the quality of high-resolution generated outputs through a multi-scale generator as well as discriminator.

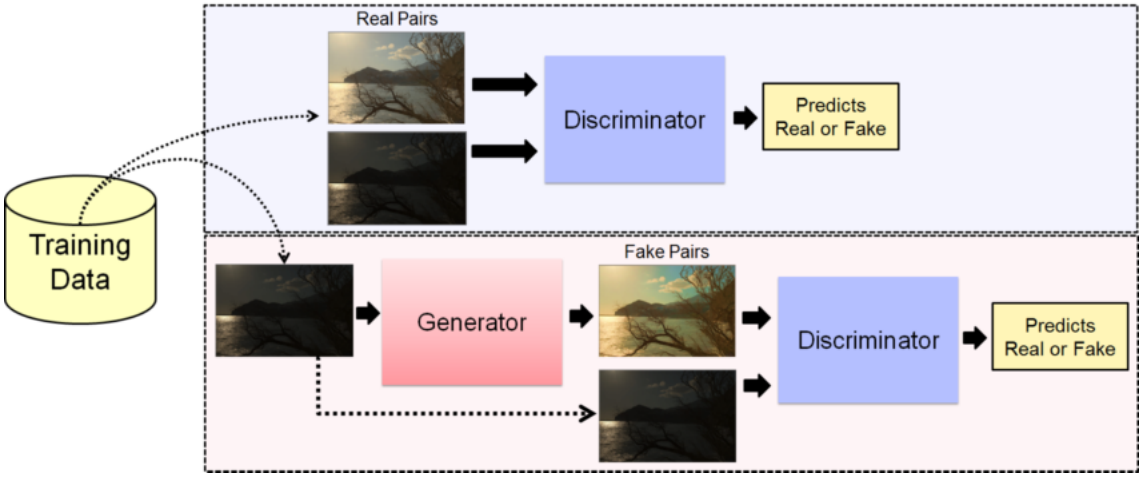
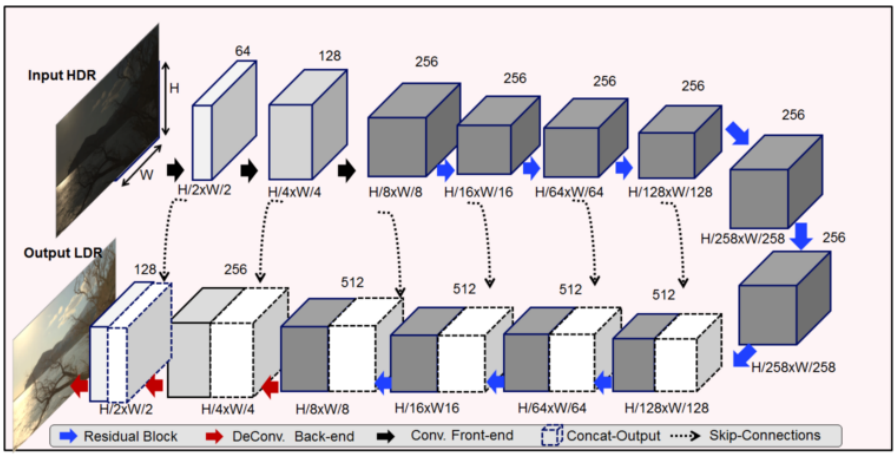


Figure 6.1 – We illustrate here the training pipeline of our Deep Tone Mapping Operator (Deep TMO). Training dataset consists of input HDRs and their corresponding best-TMQI ranked tone mapped outputs. Both the discriminator and generator are trained alternatively, first a gradient step of discriminator then of generator. While the discriminator is trained to discriminate between real and fake image pairs, the generator learns to fool the discriminator by producing synthetic tone mapped images. By doing this, the generator effectively models the underlying distribution of real ground-truth tone mapped images, thus yielding high quality results once completely trained.

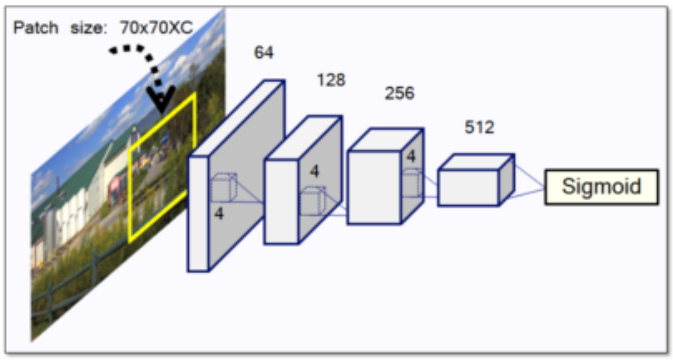
6.3 Proposed Methodology

Figure 8.15 presents an overview of our training algorithm. We train 3 deep learning based models in order to reconstruct tone-mapped LDR images from HDR images. This is effectively done by generating LDR images (fake pair in Figure 8.15) which are identical or even better than the ground-truth (GT) tone mapped images (real pair in Figure 8.15).

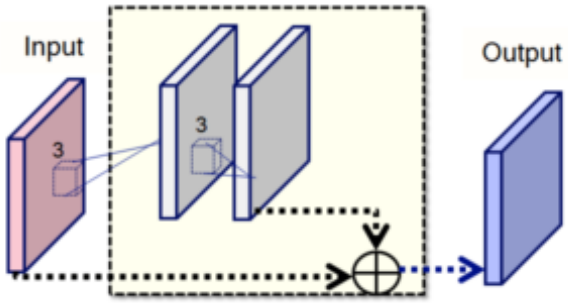
All our cGAN based tone mapping architectures have two basic modules, namely a generator and discriminator, both of which are conditioned on the *linear* HDR input. Both the generator and the discriminator compete with each other, the generator trying to fool the discriminator by producing high quality real looking tone mapped images for the given input HDR, while the discriminator trying to discriminate between real and synthetically generated HDR-LDR image pairs. Our basic discriminator architecture is similar to a PatchGAN [60, 62] which classifies patches over the entire image and averages over all of them. Similarly, the basic generator architecture comprises of an encoder-decoder network where the input HDR is given first to an encoder to yield a compressed representation which is then passed to the decoder finally resulting in a tone mapped image. Our three subsequent architectures are variants of this basic framework where DeepTMO-R is a baseline. DeepTMO-S adds skip connections between the encoder and decoder, thus effectively shuttling low level information at the time of prediction. And DeepTMO-HD constructs a multi-scale generator discriminator network, which helps in predicting tone mapped images which are both structurally consistent with the input HDR



(a) Generator Architecture with and without skip connections.

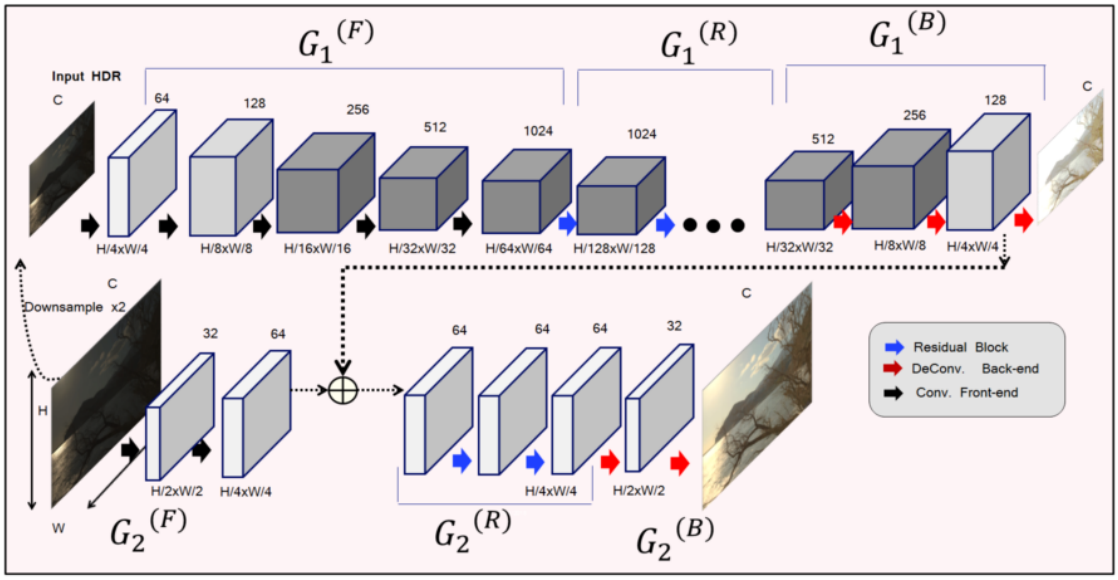


(b) Discriminator Architecture.



(c) Residual Blocks

Figure 6.2 – Here we show detailed architecture of both the discriminator and generator of DeepTMO-R and DeepTMO-S. The only difference for DeepTMO-S is the addition of skipped connections in the case of generator. The generator is framed as an encoder-decoder architecture, where the input HDR image is first passed to an encoder, which subsequently down-samples it to a compact representation. This representation is then forwarded through the decoder which up-samples it to the size of the input HDR. While the encoder consists of Convolution front end component $G^{(F)}$ and first five residual blocks $G^{(R)}$, the decoder is composed of next four residual blocks $G^{(R)}$ and a deconvolution component $G^{(B)}$. Residual Blocks consist of two sequential convolution layers applied to the input, producing a residual correction that is in turn added to the input to yield the final output. Discriminator consists of a patchGAN [48, 60, 62] architecture which is applied patch wise on the concatenated input HDR and tone mapped LDR pairs. The final prediction is an average of all the patches over the image.



(a) Generator Architecture for DeepTMO-HD is modified version of the Generator architecture of DeepTMO-R as shown in Figure 8.17a. DeepTMO-HD generator is basically a form of coarse-to-fine generator. While the finer Generator G_2 has original image as its input, input to G_1 is a $2\times$ down sampled version. This down sampled image is then effectively passed through subsequent components $G_1^{(F)}$, $G_1^{(R)}$ and $G_1^{(B)}$ which are similar to that of generator in DeepTMO-R. The final prediction from the back end G_1^B is then concatenated with the front end output of the finer-scale generator G_2. This is then passed through the back end component $G_2^{(B)}$ to yield a tone mapped output. Thus effectively our model utilizes both the coarser and finer scale information to make a prediction which results in better retaining of overall structure and minute low level details. Our discriminator architecture has identical architecture only that we give to it two different scales of input, the original and its $2\times$ down sampled version. This forces the coarse-to-fine generator to take care of both global and local details.

and at the same time preserves fine grained information recovered at different scales over the entire image.

To build our GT dataset, we provide the details in Section 6.6. All our model architectures are based upon cGAN [72] which implicitly learns a mapping from observed HDR image x and random noise vector z, to tone mapped LDR image y, given as: $G : x, z \longrightarrow y$. In the following, we discuss all the 3 different architectural models in details.

6.3.1 DeepTMO-R

Inspired from [48], our first architecture is composed of two fundamental building blocks namely a discriminator (D) and a generator (G). As shown in Figure 8.17a, input to G consists of an $H \times W \times C$ size luminance channel of an HDR image normalized between [0,1] where $C = 1$. While during training we set $H = W = 512$, inference can be performed with any larger size input. The output from G is a tone mapped image of similar size as the input. D on the other hand, takes as input pairs of luminance channels of HDR and tone mapped LDR images, and predicts whether they are real tone mapped images or fake. Thus in a way, both G and D compete with each other, G trying hard to produce outputs which cannot be distinguished from real image pairs, while the adversary D trying to detect 'fake' image pairs produced by G. Next, we discuss the architectures for both G and D which are our adaptation from [51, 120] which show impressive results for style transfer and super-resolution tasks.

Generator Architecture Our Generator is formulated as an encoder-decoder architecture as shown in Figure. 8.17a. Overall, it consists of a sequence of 3 components: convolution front end $G^{(F)}$, a set of residual blocks $G^{(R)}$ and deconvolution back end $G^{(B)}$. $G^{(F)}$ consists of 3 different convolution layers which perform subsequent down-sampling operation on their respective inputs. $G^{(R)}$ is composed of 9 different residual blocks each having 2 convolution layers, while $G^{(B)}$ consists of 3 convolution layers each of which up-samples its input by a factor of 2. For subsequent details, please see 6.4.1.

Discriminator Architecture As shown in Figure 8.17b, our discriminator architecture resembles a 70×70 PatchGAN [48, 60, 62] model, which aims to predict whether 70×70 overlapping image patches are real or fake. Such a patch-level GAN discriminator, models the high-frequency information by simply restricting its focus upon the structure in local image regions. Moreover, it contains smaller number of parameters compared to a full-image size discriminator, and hence can be easily used for any-size images in a fully convolutional manner. This discriminator is run across the entire image, and all the responses over various patches are averaged out to yield the final prediction. Further details can be found in 6.4.2

Although the generator architecture in this case yields high quality results however it still lacks the ability to reconstruct precisely local low level information as is highlighted

in Figure 6.4. Thus based upon the assumption that the underlying structure between an HDR and Tone mapped LDR image remains intact, we try to improve upon the current architecture by shuttling directly the low level information from the encoder layers to the decoder layer outputs.

6.3.2 DeepTMO-S

Various past HDR reconstruction methods, have used skip connections [93] for generating HDR scenes from single exposure [30] or multi-exposure [32] LDR images. The basic idea had been that since, both LDR and HDR scenes are different renderings of the same underlying structure, at a particular scale, their structures are also more or less aligned. Hence, it is possible to effectively transmit low level details from input to output scenes, circumventing the bottleneck of the encoder-decoder architecture. As shown in Figure 8.17a, we modify the generator, by adding skip connections between each layer i and layer n-i, n being the total number of encoder-decoder layers, which as a result concatenates all the channels at layer i with layer n-i.

As seen in Figure 6.4, Skip connections are quite helpful in retaining fine structural details and preserve contrast better, thus yielding high-quality results. However in certain cases, the quality of generated images are still unsatisfactory with certain checkerboard artifacts (see Figure 6.5). This necessitates a generator and discriminator architecture that caters to both the finer details as well as the high level semantics to generate the final tone-mapped image.

6.3.3 DeepTMO-HD

While generating high resolution tone-mapped images, it is quite evident that we need to pay attention to both low level finer details as well as high level semantic information. To this end, motivated from [111], we propose a new architecture that outlines a coarse-to-fine generator and a multi-scale discriminator in the algorithmic pipeline. We next propose further details regarding both of these architectures.

Coarse-to-Fine Generator As shown in Figure 8.18a, our Coarse-to-Fine Generator consists of two sub-architectures, namely, a Global Generator network G_1 and a local enhancer network G_2.

The architecture for G_1 is similar to G with the components, convolutional front end, set of residual blocks and convolutional back end being represented respectively as: $G_1^{(F)}, G_1^{(R)}, G_1^{(B)}$. Only difference is in the number of convolution layers and their output channels for each of the components which are clearly represented in the Figure 8.18a. G_2 is also composed of similar three components given by: $G_2^{(F)}$, $G_2^{(R)}$ and $G_2^{(B)}$ (refer to **??** for additional details).

As illustrated in Figure 8.18a, at the time of inference, while the input to a local enhancer network is a high resolution HDR image (2048×1024), the global generator receives a $2\times$ down sampled version of the same input. The local enhancer network, effectively makes tone-mapped predictions, paying attention to local fine-grained details (due to its limited receptive field on a high resolution HDR input). At the same time, it also imbibes from the global generator, a more coarser prediction (as its receptive field has much broader view). Thus the final generated output from $G_2^{(B)}$ encompasses local low-level information and global structured details together in the same tone-mapped output. Hence we get a much more structurally preserved and finely refined output as can be visualized in Figures 6.4 and 6.5.

Multi-scale Discriminator Classifying the high resolution tone-mapped output as being real or fake is a big hurdle for the discriminator too. This can be easily solved by using larger receptive field (with a deeper network) or larger convolution kernels. However it would in turn require higher memory demand, which is already a constrain while training high resolution HDR images. We basically retain the same network architecture for the discriminator as used previously, but apply it on two different scales of input calling it a Multi-scale Discriminator. We hereafter refer these two discriminators as D_1 and D_2. Inputs to the discriminators D_1 and D_2 are the original outputs generated from $G_2^{(B)}$ and a $2\times$ down sampled version respectively. The discriminators are then trained together to discriminate between real and synthetically generated images. D_2, through working on a coarser scale of image, has a larger receptive field thus having a global view of the image. This feature aids the Generator to generate more globally consistent images. D_1 on the other hand, operating at a finer scale, aids in generating more precise finer details.

6.3.4 Tone-Mapping Objective Function

The ultimate goal of our generator G is to convert high resolution HDR inputs to tone mapped LDR images, while the discriminator D aims to distinguish real tone-mapped images from the ones synthesized by the generator. We train the entire generator-discriminator architecture in a fully supervised setting. For training, we give a set of pairs of corresponding images $\{(x_i, y_i)\}$, where x_i is the luminance channel of HDR input image while y_i is the luminance output of the corresponding tone-mapped LDR image. Note that to obtain the final color output we use the used the simple ratios as used in all the previous TMO [35, 90]. Underneath we define the objective functions to train DeepTMO-R and DeepTMO-S as well as a modified variant for DeepTMO-HD.

DeepTMO-R and DeepTMO-S

The basic principle behind Conditional-GANs [72] is to model the conditional distribution of real tone-mapped images given an input HDR via the following objective:

$$\mathcal{L}_{cGAN}(G, D) = E_{x,y}[\log D(x, y)] + E_{x,z}[\log(1 - D(x, G(x, z)))] \tag{6.1}$$

where G and D compete with each other; G trying to minimize this objective against its adversary D, which tries to maximize it, i.e. $G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D)$.

In order for G to make predicted tone-mapped output being closer to real tone-mapped images, we additionally add a regularization term in the form of $L1$ distance between them which is given as:

$$\mathcal{L}_{L1}(G) = E_{x,y,z}[||y - G(x, z)||_1] \tag{6.2}$$

Thus the final objective function for both DeepTMO-R and DeepTMO-S can be defined as:

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G) \tag{6.3}$$

where λ is a regularization coefficient set to 100.

DeepTMO-HD

Since our DeepTMO-HD consists of two discriminator networks D_1 and D_2, our objective for the DeepTMO-HD architecture is:

$$G^* = \arg\min_G \max_{D_1, D2} \sum_{k=1,2} \mathcal{L}_{cGAN}(G, D_k) \tag{6.4}$$

In order to stabilize the Generator training and constrain it to yield natural image statistics at multiple scales, we append to the existing cGAN loss, an additional feature matching loss $\mathcal{L}_{FM}(G, D_k)$ (similar to perceptual loss [26, 39, 51]),given by:

$$\mathcal{L}_{FM}(G, D_k) = \mathbb{E}_{(x,y)} \sum_{i=1}^{T} \frac{1}{N_i} [||D_k^{(i)}(x, y) - D_k^i(x, G(x))||_1], \tag{6.5}$$

where D_k^i is the i^{th} layer feature extractor of discriminator D_k (from input to the i^{th} layer of D_k), T is the total number of layers and N_i denotes the number of elements in each layer. In short, we extract features from different layers of each of the discriminator and match these intermediate representations from real and generated images. Apart from this we also append a perceptual loss L_P as is used in [51] to our objective which has shown to

further improve the results:

$$\mathcal{L}_{L_P}(G) = \sum_{i=1}^{N} \frac{1}{M_i}[||F^{(i)}(y) - F^{(i)}(G(x))||_1] \tag{6.6}$$

where $F^{(i)}$ denotes the i^{th} layer with M_i elements of the VGG network [98].

Henceforth, our final objective function for a DeepTMO-HD can be written as:

$$G^* = \arg\min_{G} \max_{D_1,D2} \sum_{k=1,2} \mathcal{L}_{cGAN}(G, D) + \beta \sum_{k=1,2} \mathcal{L}_{FM}(G, D_k) + B \tag{6.7}$$

where $B = \gamma \mathcal{L}_{L_P}(G)$. β and γ controls the importance of $\mathcal{L}_{FM}$ and $\mathcal{L}_{L_P}$ with respect to $\mathcal{L}_{cGAN}$ and both are set to 10.

6.4 DeepTMO-R Architecture

In this section we specify, detailed architectural details of DeepTMO-R including the generator and discriminator.

6.4.1 Generator Architecture

$G^{(F)}$ has first a convolution layer consisting of 64 filters kernels (or output channels) each of size 7×7 applied with a stride of (1,1) and padding (0,0). Next two convolution layers with 128 and 256 filter kernels respectively and each with a size 3×3 and stride (2,2) and padding (1,1). Each of these three layers are followed by batch norm with batch size = 1 (also called instance normalization [108]) and Relu [74]. Following this, we have $G^{(R)}$ which is a set of 9 residual blocks (as shown in figure 6.2c, each of which contains two 3×3 convolutional layers, both with 256 filter kernels. Next, for $G^{(B)}$ we have two de-convolutional or transposed convolution layers with 128 and 64 filter kernels, each having a filter size of 3×3 and fraction strides of $\frac{1}{2}$. Both the layers have instance normalization and Relu added after the convolution. Finally we have another convolution layer of size 7×7 and stride 1 followed by a tanh activation function at the end. We additionally add some noise to the generator by putting dropout [100] layers in each of the residual blocks.

6.4.2 Discriminator Architecture

Discriminator architecture consists of 4 convolution layers of sizes 4×4 and stride (2,2). From first to the last, the number of filter kernels is 64, 128, 256 and 512 respectively. Each of the convolutional layer is appended with an instance normalization (except the first layer) and then leaky ReLU [66] activation function (with slope 0.2). Finally we apply a convolutional layer at the end to yield a 1 dimensional output which is followed by a sigmoid function.

6.5 Training and Implementation Details

The training paradigm for DeepTMO is inspired by the conventional GAN approach[43], where alternate stochastic gradient descent (SGD) steps are taken for discriminator (D) followed by the Generator (G). For both D and G, all the weights corresponding to convolution layers are initialized using zero mean Gaussian noise with a standard deviation of 0.02 while the biases are set to 0. Drawing from the efficacy of instance normalization over image generation tasks [108], we apply batch normalization [47] using a batch size equal to 1. All the instance normalization layers are initialized using gaussian noise with mean 1 and 0.02 standard deviation.

All our training experiments are performed using Pytorch [78] deep learning library with mini-batch stochastic gradient descent (SGD) where batch size is set to 4. We utilize an ADAM solver [54] whose initial learning rate is fixed at 2×10^{-4} for first 100 epochs and then, allowed to decay linearly to 0 until the final epoch. Momentum term β_1 is fixed at 0.5 for all the epochs. To add noise z to different layer in the generator, we apply a dropout [100] of 50 % for each convolution layer.

We also employ random jitters by first resizing the original image to 700×1100 and then randomly cropping to size 512×512. An additional mirroring is also performed before passing it through the network. All our networks are trained from scratch. Training is done on a 12 Gb NVIDIA Titan-X GPU for 1000 epochs and takes roughly 1-3 days depending upon the architecture. Inference is performed on test images of size 1024×2048 and takes less than a second.

For all the other handcrafted TMOs, we used the MATLAB-based HDR Toolbox [7] and Luminance HDR software [1]. For each tone mapping operator, we enable the default parametric setting as suggested by the respective authors.

6.6 Building HDR Dataset

In order to design a deep CNN based TMO, it is essential to obtain a large amount of dataset with a wide diversity of real-world scenes and cameras. We overcome this challenge by gathering several publicly available HDR datasets. For training the network, a total of 698 images are collected from various different sources [1, 6, 23, 24, 33, 37, 57, 80, 81, 88, 113]. From the HDR video datasets from aforementioned sources, we select the frames manually so that no two chosen HDR images are similar. All these HDR images have been captured from diverse sources which is beneficial for our objective i.e., learning a general TMO. To further strengthen the training, we apply several data augmentation techniques such as random cropping and flipping. We consider 105 images from [34] for testing purposes.

[1] http://qtpfsgui.sourceforge.net/

Target Tone Mapped Images Different TMOs results in different output tone mapped images, and consequently, this raises an essential question about which is the best tone mapped image for the input HDR scene. Several subjective studies [7] built on different hypothesis attempt to answer this question on small database of maximum 15-20 scenes. However, these solutions are impractical for large scale learning tasks. For training our Deep TMO models, a best tone mapped image is required as a target for each HDR scene. Due to unavailability of such GT images, we build the target set by using an objective Metric TMQI [114]. TMQI is a state-of-the-art objective metric which assess the quality of images on 1) structural fidelity which is a multi scale analysis of the signals, and 2) naturalness, which is derived using the natural image statistics. To find the GT for each training image, 13 classical TMO are considered [7] which includes [4, 21, 27, 29, 35, 58, 67, 79, 90, 95, 105] and gamma and log mappings [7]. Then, based on the TMQI scores, the best tone mapped output for an individual scene is selected. The selection of these tone mappings is inspired from the following subjective evaluation studies [13, 59] which highlight the distinctive characteristics of mapping functions, which we aim to inculcate in the learning of our Deep TMOs.

6.7 Results and Evaluation

In this section, we present an overview of our tone mapping models in terms of visual quality and quantitative performance. We first discuss the specific characteristics of three proposed models, including their adaptation to display, scene content, and sharpness in rendering high-resolution tone mapped outputs. Later we compare the performance of our best quality results with several handcrafted tone mapping methods [4, 21, 27, 29, 67, 79, 90, 95] on 105 images of test dataset [34]. Note that these scenes are different from training set and have not been seen by any of the architectures while training.

6.7.1 Comparison of the Three Architectures

We compare our three architectures in Figures 6.4 and 6.5, where in each column we showcase the full tone-mapped image for each of the architectures together with a cropped inset. In Figure 6.4, all the architectures give over all good quality LDR results with some subtle differences. From the cropped insets in second row, both DeepTMO-S and DeepTMO-HD retain the texture on background wall, while DeepTMO-R slightly blurs them out. Same goes for the bottommost window panels. This is quite evident, due to the lack of skip connections in its architecture (Figure 8.17), that help in transferring information from encoder to the decoder layers for better reproduction of output. While comparing between DeepTMO-HD and DeepTMO-S, we see DeepTMO-HD is able to better preserve very minute details such as frames in lower window in the inset. This is largely due to the presence of its two scale generator which caters to both finer details and

coarser structures from the high-contrast linear HDR maps while generating tone-mapped image.

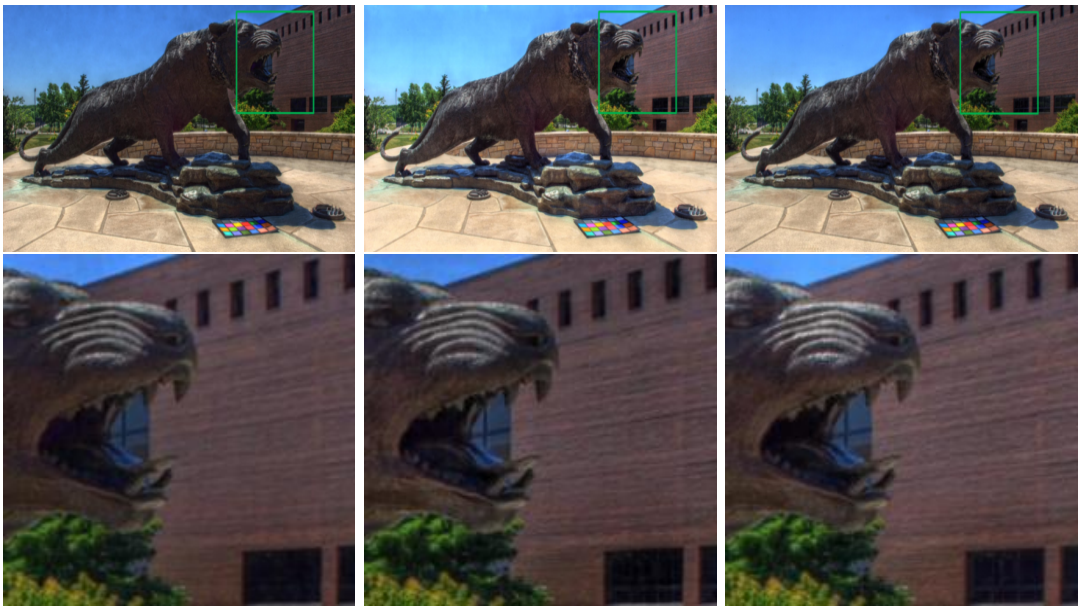
Next, in Figure 6.5 we show another interesting high-contrast natural scene where the three architecture provide similar contrast images but with some prominent visible effects. From the cropped insets, we see that DeepTMO-R results in blurry effect on the textured bark of tree, similar to previous example. DeepTMO-S on the other hand, doesn't produce any blurriness, but instead, we notice pronounced repetitive checkerboard artifacts. Such artifacts have been recently discussed in deep-learning based image rendering problems [38, 76] and are mainly caused due to no direct relationship among intermediate feature maps generated in de-convolutional layers. Nevertheless, it is still an open problem. DeepTMO-HD, on the other hand, gives us sharper and checkerboard free images while preserving the fine-details too. One possible reasoning behind this can be that the discriminator part working at the finer scale, effectively distinguishes a fake image through these artifacts, and thus forces the generator to subsequently generate checkerboard free images.

Overall, we observe that the result produced by our models are aesthetically pleasing images with a good balance of color contrast and details. Even though we selected GT images based on TMQI which is a color blind metric, our generator could develop an implicit understanding of contrast and saturation through the underlying data distribution and reproduce such impressive colors in the output images. Unlike past CNN based HDR techniques, which target to minimize Euclidean distance between the GT and output, our model adapts the cost based upon the target dataset and is effectively able to hallucinate high quality tone mapped results. Amongst the three models, our DeepTMO-HD successfully maps the linear-input HDR content in the most refined manner, preserving the minute details and the overall structural content. We henceforth use DeepTMO-HD for all the subsequent comparisons.

6.7.2 Comparison with TMOs

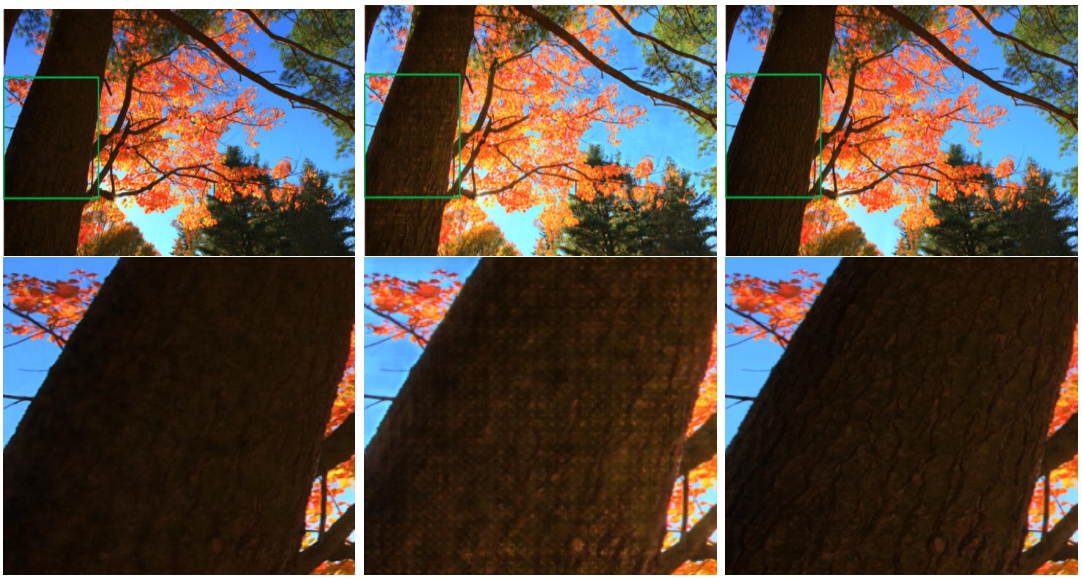
We begin the comparison of DeepTMO model against the Best Ranked Tone Mapper (BRTM) test scenes to assess the overall generalization capability. To obtain the best ranking tone mapped test scene, we follow a similar paradigm as provided in Section 6.6. Later, one-to-one comparison are drawn between the existing TMOs and the proposed tone mapping model.

Comparison with Best Ranked Tone Mapped Scenes Figure 6.6 demonstrates a set of HDR images from the test dataset that have been mapped using the DeepTMO-HD and corresponding BRTMs. These sample scenes depict the exemplary mapping of the linear HDR content using DeepTMO-HD where it successfully generalizes over variety of scenes and even outperforms the respective BTRM in terms of overall quality of contrast preservation and visual appeal. In the first row, while the BRTM which is DurandTMO in



DeepTMO-R DeepTMO-S DeepTMO-HD

Figure 6.4 – Compared between the three proposed architecture DeepTMO-R, DeepTMO-S, DeepTMO-HD (I). From the insets, DeepTMO-R suffers from blurriness issues in the wall and lowermost window panels. DeepTMO-HD and DeepTMO-S both are able to preserve the finer details, though the window panels are much more clearly visible in case of DeepTMO-HD



(a) DeepTMO-R (b) DeepTMO-S (c) DeepTMO-HD

Figure 6.5 – Comparisons between three architectures (II). As seen in the inset, while DeepTMO-R simply results in blurred outputs in the bark of tree, DeepTMO-S tries to refine them but is faced by *checkerboard* artifacts [38, 76]. The DeepTMO-HD provides best results amongst the three methods while preserving the fine details, contrast and sharpness in the image.

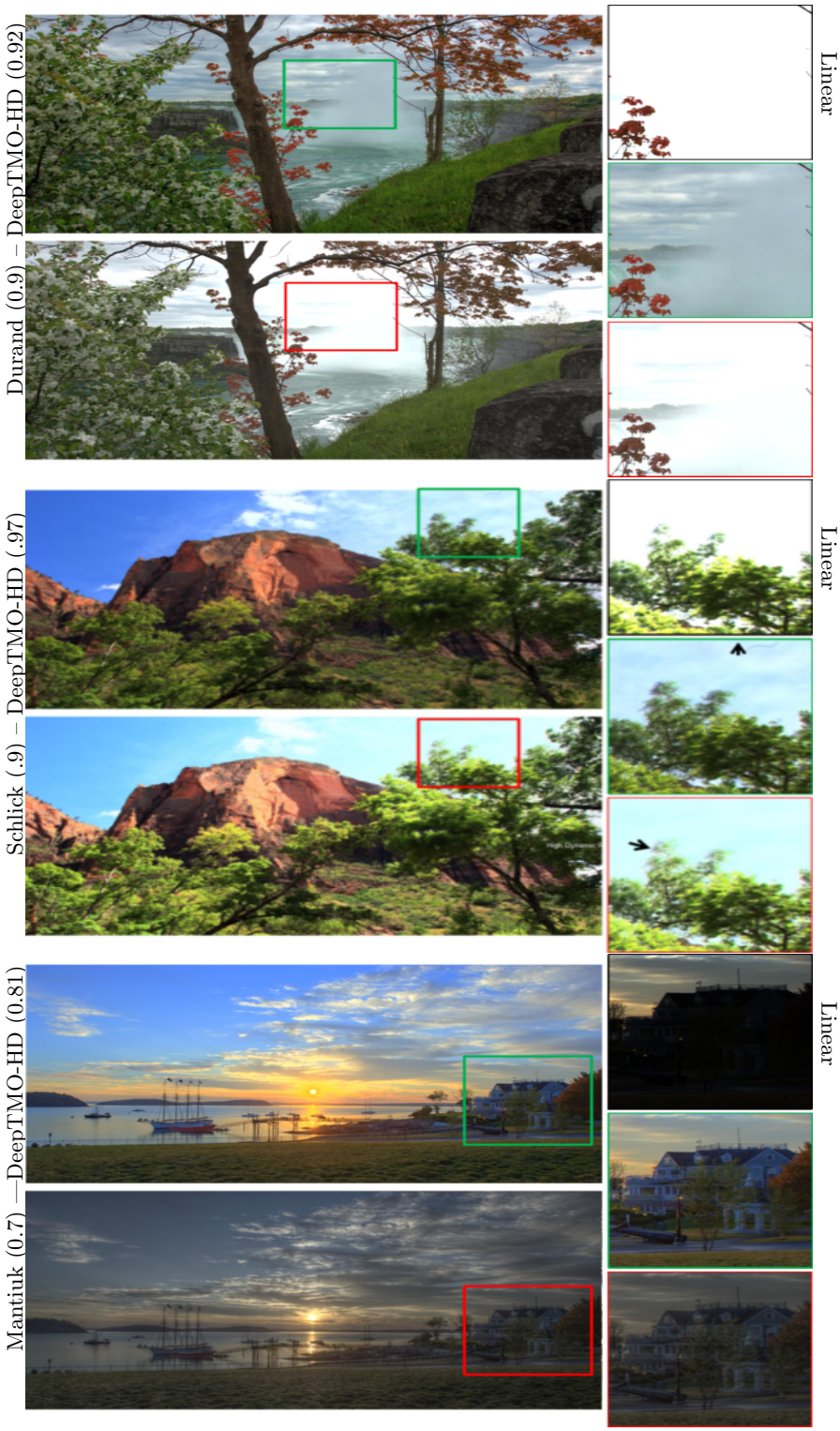


Figure 6.6 – Comparison of our method with the respective BestTMOs based on the TMQI scores with the highlighted zones for scene: The Canadian Falls (row I), The Grotto (row II) and the Bar Harbor Sunrise (row III). Zoom-ins for each scene highlights mapping outputs for DeepTMO-HD and the respective BestTMO for the corresponding HDR-linear input. We notice that our model has no saturation effect in waterfall (row I), preserves finer details in sky (row II) and effectively balances luminance for the house (row III). The TMQI scores for each scene are provided alongside each TMO.

this case fails to recover the overly saturated regions, our model successfully preserves the fine details in the sky along with the waterfall and mountains in the background. Similarly in row 2, while SchlickTMO fails to preserve saturated regions in sky, our model is able to preserve both textures of the clouds as well as the bird in sky. We additionally are able to retain textures of the leaves which again for Schlick get partly saturated. Same goes for a dark scene in third row, where our trained DeepTMO-HD generator compensates the lighting and preserves the overall contrast of the generated scene which isn't the case for MantiukTMO. Hence we see that, learned with different mapping characteristics, our multi-scale generator network accordingly adapts for each scene to provide an image of convincing quality, which was missing in all the existing hand-crafted tone mapping methods.

To further demonstrate the generalization capability of DeepTMO models on all the 105 real world scenes, in Figure 8.19, we show a distribution plot of the number of scenes against the TMQI Scores. For completeness, we also add scores achieved by BRTMs. The curves clearly show that the generated tone mapped images for all models compete closely with the best available tone mapped images on the TMQI metrics with DeepTMO-H fairing the best among all. Hence, consolidating our main motivation of using this TMO.

Comparison with TMOs We first provide quantitative analysis in Table 8.2, to compare the performance of our proposed model with the existing approaches. For each method, the TMQI scores are averaged over 105 scenes of the test dataset. Final averaged results show that all our three proposed tone mapping model outperform all the existing methods, hence proving their generalization capability over the rest. Moreover, while all three proposed models compete closely, our DeepTMO-HD model performs slightly worse than the DeepTMO-S. This is perhaps because the skip-connections provide marginal gains in some of the scenes. However, we believe that multi-scale DeepTMO-HD architecture is slightly more stable and produces visually pleasing results, and hence, we use it in the rest of the paper for comparisons.

Next, in Figure 8.20, we demonstrate qualitative comparisons of our model with four existing methods namely, Mantiuk [67], Reinhard [90], Fattal [35] and Durand [29] over 5 real world scenes from the test dataset. For clarity of results, we choose only top four methods based on the scores in Table 8.2. From all the five scenes, we observe that our proposed tone mapping adapts to the content and preserves the contrast and fine details in the most convincing way in comparison to all the other local and perceptual TMOs. While perceptual methods such as [67] gives well balanced fine detailed information about the scenes, the resulting outputs are somewhat washed out in terms of overall perceptual brightness and visual appeal. Same goes for [29] for scene 1 and scene 2. The approach in [90] loose on preserving details in the overly saturated regions *e.g.*, clouds in scene 1. Although in these five scenes, approach in [35] provides better quality results as compared to other hand-crafted methods, in scene 1 and 5 it slightly lags behind our DeepTMO-HD

and provides darker results. Altogether, we observe that our DeepTMO-HD addresses the generality concept and adapts to a variety of HDR content. Based on its high level understanding of the content in the scene, it efficiently recovers convincing colors, structure and the finer details of the scene.

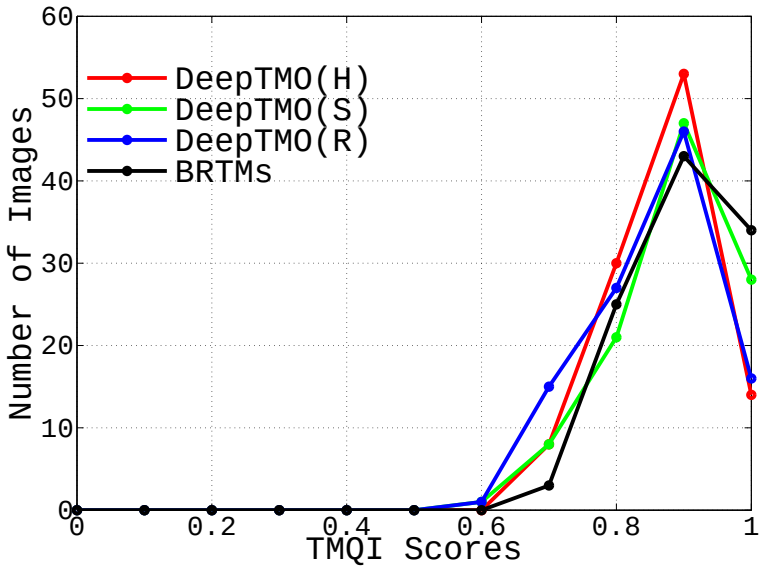


Figure 6.7 – Quantitative performance comparison of DeepTMO-R, DeepTMO-S and DeepTMO-HD with the BestTMOs.

6.7.3 Limitations

Though our model successfully demonstrates generality in addressing wide variety of scenes, its expressive power is limited by the amount of available training data and quality of its corresponding ground truths. Due to unavailability of subjectively annotated 'best tone mapped images' for the HDR training dataset, we resort to an objective TMQI metrics to build the corresponding ground truth LDR. However, the metric itself is not as perfect as the human visual system. We illustrate this point in figure 6.9. The images ranked lower by TMQI metric in column 3 and 4 are somehow more interesting than their best-ranked counterpart in column 2. For *e.g.*, in row 1, while the best ranked TMO outputs bursty effects near the lamp and doesn't effectively preserves the textual details in the book, TMOs in column 3 and 4 preserves better structural details and naturalness. Similarly in row2 for the best TMO, we can visualize some overall hazy effects as well as some noisy artifacts on the front board, which isn't the case in other two TMO's.

Such samples can eventually restricts the generation power of our model for darker scenes, specifically in high illumination regions such as lamps or bulbs. The discriminator no longer forces the generator to eradicate such bursty effects as the underlying ground truth samples, with which it is trained, are noisy. To overcome these challenges we provide

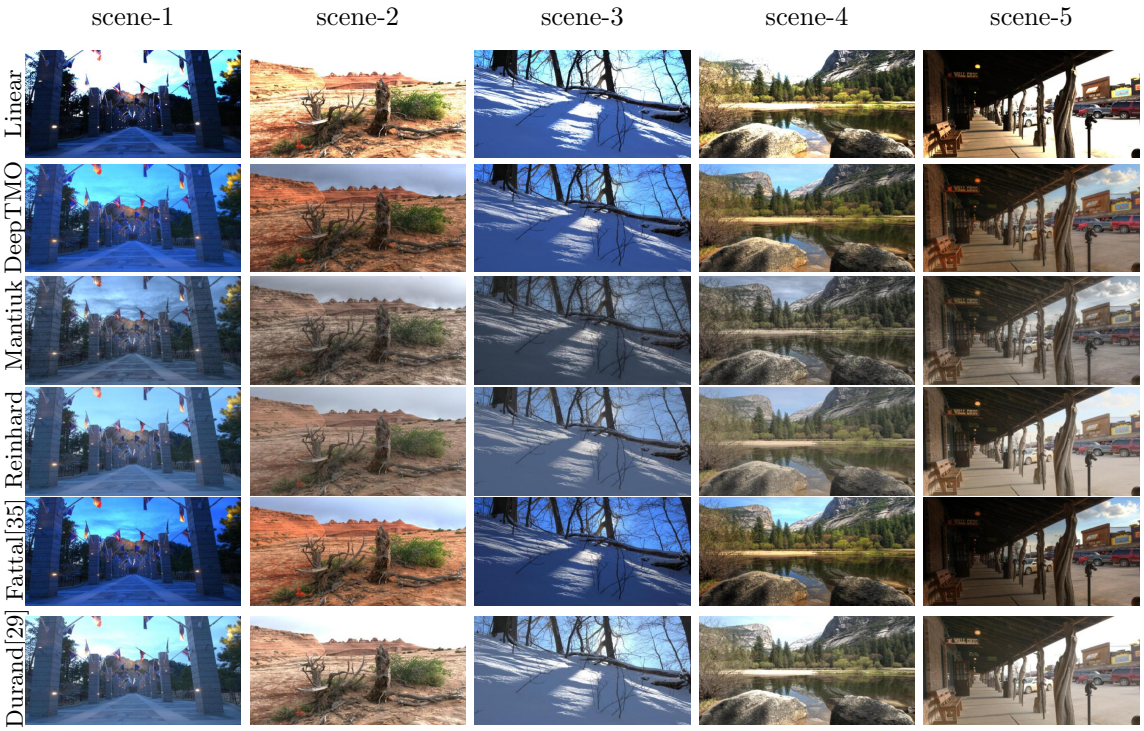


Figure 6.8 – *Qualitative Results*. Five sample scenes from *Fairchild HDR dataset*, taken with different natural lighting variations.

Table 6.1 – *Quantitative Results*. mean TMQI scores on the test-set of 105 images from Fairchild HDR database.

TMOs	TMQI
Ward [58] TMO	0.71 ±0.07
Pattnaik [79] TMO	0.78 ±0.04
Log [7] TMO	0.72 ±0.09
Gamma [7] TMO	0.76 ±0.07
Ashikh [4] TMO	0.70 ±0.06
Durand [29] TMO	0.81 ±0.10
Tumblin [105] TMO	0.69 ±0.06
Drago [27] TMO	0.81 ±0.06
Schlick [95] TMO	0.79 ±0.09
Reinh [90] TMO	0.84 ±0.07
Fattal [35] TMO	0.81 ±0.07
Chiu [21] TMO	0.70 ±0.05
Mantiuk [67] TMO	0.84 ±0.06
DeepTMO-HD TMO	0.87 ±0.06
DeepTMO-S TMO	0.88 ±0.07
DeepTMO-R TMO	0.86 ±0.08

DeepTMO-HD BRTM BRTM-Rank2 BRTM-Rank3

Figure 6.9 – Sample cases where top scoring TMQI's TMO shows not-so-visually desirable outputs. In column I, we have tone mapped outputs from DeepTMO-H, in column II for BRTM , while in column III and column IV we provide results for two other top ranking TMO's.

some possibilities in Chapter 7.

6.8 Conclusion

In this chapter, we have presented three different deep tone mapping architectures for a perceptual objective. Tailored in a generative-discriminative framework, the proposed models are trained to output realistic looking tone-mapped images, that duly encompass all the various distinctive properties of the available tone mapping operators. For completeness, we provide an extensive comparison among the three proposed frameworks highlighting the role that each design variation plays in their respective reproduced outputs.

Our deep tone mapping models also overcome the frequently addressed blurry or tiling effects in recent HDR related works [31, 32], a problem of significant interest for several high-quality learning-based graphic rendering applications as highlighted in [31]. By simply learning an HDR-to-LDR specific cost function, the proposed models successfully preserve desired output characteristics such as underlying contrast, lighting and minute details present in the input HDR at the finest scale. Lastly we validate the versatility of our methodology through a detailed quantitative and qualitative comparisons with existing tone mapping techniques.

The work presented in this chapter has been submitted to SIGGRAPH'18.

Chapter 7

Conclusions

7.1 Summary

This thesis presents the quantitative and qualitative analysis of HDR imagery for specific task measures. More specifically, this thesis explores advantages of HDR content for specific tasks and timely addresses the corresponding limitations and challenges. In the following three major aspects, we highlight the contributions made in this thesis.

Firstly, based on the limited state-of-the-art studies, we primarily identify the most natural and important questions in this direction. We begin with a performance evaluation study on what specific advantages can HDR images bring over LDR for a given feature extraction task. To this end, we propose a geometrically calibrated dataset with wide range of illumination condition. We then observe how different HDR modalities can impact feature extraction performance. Since no modality performs best across all scenes, we address the need of an optimal design to use HDR information and further, investigate the factors influencing the design essential for optimal modalities through a small experimental study.

Secondly, considering our findings from evaluation study, we propose three in-line methodologies to optimally use the HDR information to enhance the efficiency of local features extraction. By adapting a variant of Bilateral filtering, we showcase models' learning capability by brining invariance to luminance change at three levels, namely, keypoint detection, description and final matching. Finally, we evaluate the performance of all the learning-based models on a proposed HDR dataset of 8 indoor/outdoor scenes where it outperforms state-of-the-art TMOs across different feature extraction algorithms.

Thirdly, to handle a large variety of HDR real-world images, we present three end-to-end deep learning based generic tone mapping designs which caters to desired task-specific characteristics. Our previously proposed TMOs essentially requires a specific filtering design which need to be differentiable. Since not all the TMOs can satisfy this criteria, DeepTMOs overcome these drawbacks and learn from a variety of filtering characteristics while following an easy differentiability. Furthermore, DeepTMOs can serve as a baseline

for future task-optimal HDR designs as they can be fine tuned for any specific task by simple cascading.

In the following, we expand the above mentioned aspects in correspondence to chapters of the thesis.

- We investigate how much gains can HDR bring over LDR for the feature extraction stages and which are the best modalities of using HDR to obtain it. To this end, we prepared the framework with 11 HDR based modalities, 2 keypoint detection and 4 full feature extraction schemes. Additionally, we propose a geometrically calibrated HDR luminance change dataset with a variety of lighting variations.

  The analysis based on quantitative performance measures of keypoint detection and feature matching scores on different scenes confirms the potential of HDR techniques over single LDR exposures. For both detection and matching, we observe the linear HDR values are inappropriate to be used for LDR optimized visual recognition tasks. In case of TMOs, we observe that their performance varies with the type of scene, exhibiting their nature of content-dependence. Furthermore, we observe that all the local TMOs producing very appealing results are not necessarily the best option for image analysis tasks. More interestingly, we have also observed that local TMOs with very high repeatability rate for feature detection are not necessarily the best option when the full feature extraction pipeline is considered. For individual test pairs cases, we find no modality which is absolutely outperforming. Hence, it remains unclear whether HDR pixels should be encoded approximatively linear to perception or directly tone map using existing functions. Therefore, a more optimal means of modality needs to be designed. (Refer to Chapter 3)

- We develop a learning based adaptive tone mapping framework for HDR images which results in stable and efficient keypoint detection. To this end, we initially conduct an experimental study investigating what it takes to optimize a tone mapping function for a specific task such as keypoint detection. Build on the observations, we propose a new learning-based adaptive tone mapping framework which aims at enhancing keypoint stability under drastic illumination variations. To this end, we design a pixel-wise adaptive TMO which is modulated based on a model derived by Support Vector Regression (SVR) using local higher order characteristics. To circumvent the difficulty to train SVR in this context, we further propose a simple detection-similarity-maximization model to generate appropriate training samples using multiple images undergoing illumination transformations.

  We evaluate the performance of our proposed framework in terms of keypoint repeatability for state-of-the-art keypoint detectors. Our experimental results showcase the efficiency of our proposed learning-based adaptive TMO which yields higher keypoint stability when compared to existing perceptually-driven state-of-the-art

TMOs. (Refer to Chapter 8)

- The keypoint detection and description designs are independent in nature. After optimizing the TMO for keypoint detection, we address tone-mapping optimality for full feature extraction algorithms. To this end, we first propose a descriptor-optimal TMO design which solely aims at the extraction of invariant (as much as possible) descriptors from high-contrast areas of the scenes. Then, we collectively address both stages of keypoint detection and descriptor extraction in the feature matching framework. To this end, we first propose an energy maximization model to generate appropriate training samples by subsequently addressing the detection and description costs. Then, by locally altering the intrinsic characteristics of tone mapping function, guidance model is learned to predict optimal parameter-maps.

  We evaluate both proposed TMOs on a HDR dataset of indoor/outdoor scenes where they outperforms state-of-the-art TMOs across different image matching algorithms. Our proposed task-optimal TMOs showcase their versatility when applied to different detection/description approaches and hence, can be directly plugged into various local feature based applications. (Refer to Chapter 5)

- We propose deep learning based TMOs which predict high quality tone mapped outputs over a wide spectrum of *linear* HDR images. Currently, the end-to-end deep learning based TMOs are trained with a perceptual objective to yield most natural images, thanks to the large availability of HDR images. Due to ease in backpropagation, the proposed model can be simply fined tuned with any desired objective such as image matching. Therefore, it eradicates the need of designing any proxy cost functions. Additionally, these models can serve as a baseline architecture to explore HDR imagery for several other domain specific analysis tasks such as medical image analysis or high resolution remote sensing tasks.

  Based on conditional generative adversarial network (cGAN), our proposed architectures learn to adapt to a wide variety of content and tackle HDR-specific challenges such as contrast, brightness and luminance, while preserving fine details. Aiming for high quality LDR, we address some prominent issues like blurring, tiling patterns and saturation artifacts encountered in past HDR related deep learning methods. To this end, we propose an additional high resolution prediction mechanism which caters to such finer details. To further leverage on the large availability of unlabeled high dynamic range data to train our network, we rely on an objective HDR quality metrics called Tone Mapping Image quality Index or TMQI. Finally, we demonstrate that our proposed deep tone mapping models generate high quality realistic output images and outperform all other classical tone mappings to generalize well over a larger spectrum of real-world scenes.(Refer to Chapter 6)

7.2 Future Research Directions

Wide scale availability of HDR images and videos databases have opened up new analytical perspectives for future research. In this concluding section, we discuss several possible extensions of this thesis.

Investigation of HDR imagery for Dynamic Scenes

As evident from the results from Figures 4.13, 8.14, information preserved by HDR images facilitates the extraction of highly stable and luminance invariant local features. The tested scenarios consider HDR images that are taken from a *static scenes*. However, real-time scenarios can be dynamic and consequently more challenging. This is mainly due to different combinations of physical transformations such as geometric (rotation, viewpoint change), deformational variations and due to sensor noise. One practical scenario includes lighting + viewpoint changes with mobile platforms such as drones.

The presented learned models theoretically should adapt in accordance to the invariance property of the feature extraction algorithms, as shown in Figure 5.20 for in planar rotation. However, for out-of-the plane rotation problems that are specific for a mobile capturing platform, our models need to be practically re-calibrated. Note that no state-of-the-art local feature extraction algorithms is best under all transformations [116, 119]. Therefore, instead of simply learning the regressor models with basic corner and descriptor based features, multiple characteristics from feature extraction algorithms, *e.g.* the ones evaluated in [119], needs to be infused. Additionally, it requires a proper geometrically calibrated dataset which needs to be created.

Perception Vs Vision

In Chapter 6, we proposed the deep learning based tone mapping methods for perceptual tasks. As already mentioned in Section 6.1, one possible extension is to design a task-optimal deep-learning based TMO by fine tuning the DeepTMO model. This could help us to compare the results obtained from deep networks learned from two different objectives (perceptual and computer vision) over an HDR dataset. Since HDR technology gives representation of real-world scenes closer to human eyes, the comparison between the two models will further open up the possibilities of research in deeper understanding of these networks, finding their technical explanations about inspirations drawn from human brain.

HDR Analysis in Temporal domain

A majority of this thesis explores the versatility of high dynamic range imagery in enhancing the stability of local features in RGB images. However, one natural extension is to upgrade the analysis for HDR videos. Information in temporal domain has far more potential for real-time computer vision applications such as surveillance tasks. A vast amount of

information that are lost in low-contrast scenes can be reconstructed using predictive modelling using an additional temporal dimension of the HDR videos. This could boost the performance in several video based tasks such surveillance applications, real-time tracking, analysis of gestures and actions.

Deep Learning in HDR Imagery with small datasets

In comparison with millions of LDR annotated images, publicly available HDR dataset are very small, which in turn limits the exportability of HDR technology. Though an optimal solution would be to create a large HDR dataset with subjectively evaluated ground truths, this would be quite a cumbersome and tedious task. A work around can be reverse engineering the training dataset by reconstructing their corresponding HDR using recent deep learning based inverse tone mapping models [30, 32].

Another alternative future work can be to rely on limited amount of samples, augmented with some noisy samples and then, utilizing a weakly supervised learning paradigm [65]. For tasks such as HDR-to-LDR mapping, completely unsupervised learning is also possible without giving any input-output pairs [120]. The intuition is to allow the network to decide by itself which is the best possible tone-mapped output simply by independently modeling the underlying distribution of input HDR and output tone mapped images.

Publications

Journal articles

1. A. Rana and G. Valenzise and F. Dufaux, "Learning-based tone mapping operator for efficient image matching", *IEEE Transaction of Multimedia(TMM)*, 2017 accepted.

Conference/Workshop papers

1. A. Rana and G. Valenzise and F. Dufaux, "Evaluation of Feature Detection in HDR Based Imaging Under Changes in Illumination Conditions", *IEEE International Symposium on Multimedia (ISM)*, Miami,USA, December, 2015.

2. A. Rana and G. Valenzise and F. Dufaux, "An Evaluation of HDR Image Matching under Extreme Illumination Changes", *The International Conference on Visual Communications and Image Processing (VCIP)*, Chengdu, China, 2015.

3. A. Rana and G. Valenzise and F. Dufaux, "Optimizing Tone Mapping Operators for Keypoint Detection under Illumination Changes", *2016 IEEE Workshop on Multimedia Signal Processing (MMSP 2016)*, Montréal, Canada, 2016.

4. A. Rana and G. Valenzise and F. Dufaux, "Learning-based Adaptive Tone Mapping for Keypoint Detection", *The International Conference on Visual Communications and Image Processing (ICME)*, Hong Kong, China, 2017.

5. A. Rana and G. Valenzise and F. Dufaux, "Learning-based Tone Mapping Operator for Image Matching", *IEEE International Conference on Image Processing (ICIP'2017)*, Beijing, China 2017.

6. A. Rana, P. Singh, G. Valenzise, F. Dufaux and N. Komodakis. "Deep Tone Mapping Operator for High Dynamic Range Imagery", *ACM SIGGRAPH*, 2018 submitted.

Chapter 8

Résumé de thèse

8.1 Résumé

La technologie HDR (High Dynamic Range) a acquis une immense popularité pour sa capacité à représenter une large gamme de couleurs et d'intensités lumineuses présentes dans des environnements réels [28, 68]. Dans un sens, ces images nous permettent de dessiner des détails subtils, mais discriminants, présents à la fois dans les zones extrêmement sombres et claires d'une scène, qui autrement se perdraient dans l'imagerie traditionnelle à gamme dynamique basse (LDR). Avec les récents progrès de l'intelligence artificielle, que ce soit sous la forme de voiture autonaume ou de dispositifs de surveillance automatisés, un tel contraste préservant les propriétés HDR est essentiel pour la compétence des algorithmes de vision par ordinateur sous-jacents. En d'autres termes, ces algorithmes devraient être capables d'analyser efficacement chaque région d'une scène sans trop d'incertitude.

Bien que ces algorithmes soient entièrement personnalisés pour les images LDR capturées dans des conditions différentes, ils échouent lamentablement dans les scènes à contraste élevé ayant une luminance élevée ou faible [44, 110, 117, 119]. Puisque les scènes à contraste élevé sont extrêmement courantes dans le monde réel, cela devient très critique dans des cas comme les véhicules automatisés où des vies humaines sont en danger impliqués. Ainsi, l'utilisation de la technologie HDR est néceéssaire pour la viabilité des algorithmes de vision par ordinateur. Bien qu'un grand nombre d'algorithmes aient été concus pour pour interpréter des scènes surexposées ou sous-exposées à l'aide d'images LDR, peu de travail a été fait jusqu'à présent dans le contexte du contenu HDR.

Cette thèse est centrée sur l'analyse d'images HDR "enrichies" au profit d'un problème de correspondance des caractéristiques visuelles de bas niveau, qui est la base de nombreux autres algorithmes de vision par ordinateur de haut niveau, y compris l'enregistrement et la vision stéréoscopique, le mouvement l'estimation et la localisation, l'appariement, la récupération et la reconnaissance d'objets et d'actions. Plus spécifiquement, la thèse examine les défis fondamentaux liés à l'utilisation du HDR et en déduit les moyens optimaux d'utiliser le contenu HDR pour améliorer la robustesse de telles tâches.

Cette thèse présente l'analyse quantitative et qualitative de l'imagerie HDR pour des tâches spécifiques. Cette thèse explore tout d'abord les avantages du HDR pour des tâches spécifiques, ensuite, aborde en temps opportun les limites et les défis propres à chaque tâche.

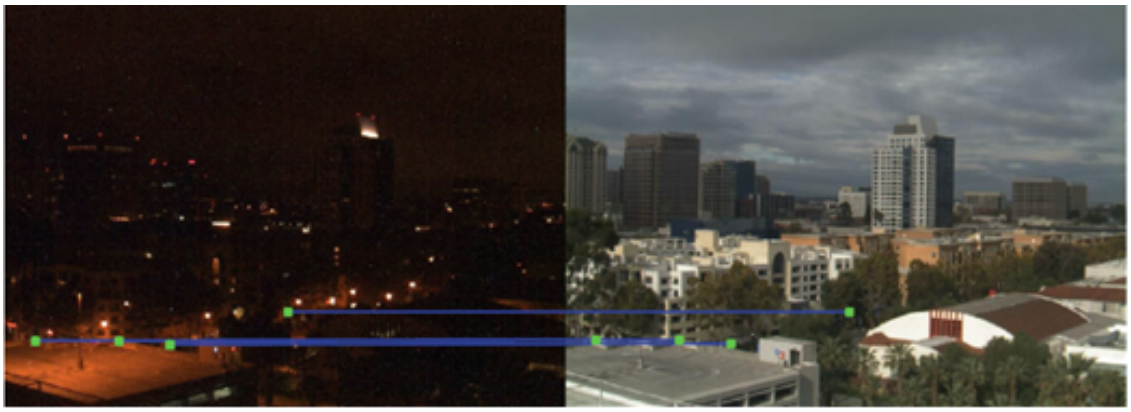
Dans cette thèse, nous soulignons les contributions apportées dans les trois aspects majeurs qui suivent.

Dans le premier aspect, nous identifions les questions les plus naturelles et les plus importantes sur la base des études limitées de l'état de l'art dans cette direction. Nous commençons par une étude d'évaluation des performances sur les avantages spécifiques que les images HDR peuvent apporter par rapport aux images LDR pour une tâche donnée d'extraction de features. A cette fin, nous proposons un jeu de données géométriquement calibré avec un large éventail de conditions d'éclairage. Nous observons ensuite comment les différentes modalités du HDR peuvent avoir un impact sur les performances d'extraction des features. Puisqu'aucune modalité n'est la plus performante sur toutes les scènes, nous devons répondre au besoin d'une conception optimale qui utilise l'information du HDR. Nous étudions plus en détail les facteurs qui influencent la conception essentielle pour des modalités optimales par le biais d'une étude expérimentale.

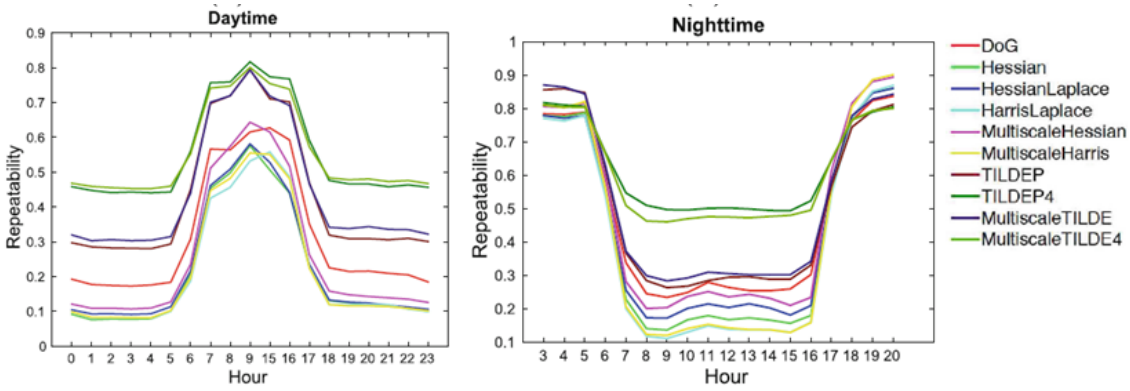
Deuxièmement, nous proposons trois méthodologies spécifiques tout en tenant compte des résultats de l'étude d'évaluation. Nous proposons d'utiliser de manière optimale les informations HDR pour améliorer l'efficacité de l'extraction des caractéristiques locales. En adaptant une variante du filtrage bilatéral, nous mettons en évidence la capacité d'apprentissage des modèles. Nous proposons d'apporter l'invariance au changement de luminance à trois niveaux : (1) la détection des points clés, (2) la description et (3) la mise en correspondance des images. Enfin, nous évaluons la performance de tous les modèles proposés en fonction de l'apprentissage, sur la base d'un ensemble de données HDR proposé de 8 scènes intérieures/extérieures. Nous montrons en outre comment notre modèle surpasse les Opertateur de Tone mapping (TMOs) les plus perfectionnés à travers différents algorithmes d'extraction de caractéristiques.

Troisièmement, pour traiter une grande variété d'images HDR du monde réel, nous présentons trois modèles génériques de cartographie tonale basés sur l'apprentissage profond 'end-to-end' qui répondent aux caractéristiques spécifiques des tâches souhaitées.

Les TMOs que nous avons proposés précédemment exigent essentiellement une conception de filtrage spécifique qui doit être différentiable. Comme tous les TMO ne peuvent pas satisfaire ces critères, DeepTMOs surmonte ces inconvénients et apprend d'une variété de caractéristiques de filtrage et qui est facilement différentiable. De plus, les DeepTMOs peuvent servir de référence pour les futures conceptions HDR optimales des tâches, car elles peuvent être affinées pour n'importe quelle tâche spécifique en simple cascade.



(a) Exemple d'appariement Jour-Nuit



(b) Taux de répétabilité de détection des points clés

Figure 8.1 – En (a), nous montrons un exemple de [119], où l'appariement des points saillants (communs aux deux images) est représenté par des lignes bleues entre deux images de la même scène prises à des heures différentes de la journée. En (b), nous montrons le taux de répétabilité de la mesure d'efficacité sur un grand ensemble de données d'images LDR jour/nuit en utilisant les techniques de pointe. La crête et les creux dans les courbes illustrent que l'image capturée pendant le jour correspond bien avec seulement d'autres images de jour et non avec celles capturées dans l'obscurité.

8.1.1 Context

La robustesse des applications de vision par ordinateur peut être interprétée à partir d'une hiérarchie à trois niveaux : bas, moyen et haut niveau. Puisque ces deux derniers niveaux se construisent à partir du premier, l'efficacité dans l'analyse de bas niveau est essentielle [116].

En général, l'analyse de bas niveau est définie et évaluée en fonction du problème de "correspondance visuelle" [96]. Le problème est formulé en dessinant la correspondance entre les images à l'aide d'algorithmes d'extraction des caractéristiques visuelles. Les caractéristiques visuelles sont les signatures discriminantes qui contiennent des informations locales provenant des emplacements saillants des images. La correspondance entre ces caractéristiques définit la "compatibilité " entre les deux contenus.

Un exemple provenant de la référence [119] est présenté en Figure 8.1 (a), illustrant la correspondance entre une scène de jour et une scène de nuit à l'aide d'images LDR. Plusieurs tentatives, y compris des modèles locaux [44, 103, 112], des modèles de normalisation globale[97] et des méthodes basées sur l'apprentissage [110, 116], ont été faites pour assurer une meilleure conception des invariants de luminance dans l'imagerie LDR. Cependant, ces techniques sont pratiquement inefficaces pour compenser complètement la perte d'information ou comprendre le changement dans la configuration spatiale des objets présents dans la scène. Par conséquent, ces algorithmes ne parviennent pas à trouver de vraies correspondances entre des objets similaires et entraînent une forte baisse de performance. Dans l'exemple d'appariement jour/nuit de la Figure 8.1, la performance de plusieurs algorithmes perfectionnés de détection de caractéristiques diminue de façon significative dans les variations d'éclairage jour/nuit.

L'imagerie HDR, d'autre part, peut partiellement surmonter ces limitations en capturant une large gamme d'éclat et de luminosité tout en préservant les détails fins dans à la fois dans les régions sombres et les région très lumineuses. Par conséquent, en raison de ces capacités étendues, l'utilisation de l'imagerie HDR dans l'extraction d'entités locales est essentielle.

Les algorithmes d'extraction d'entités locales ont été largement explorés dans la littérature sur la vision par ordinateur. Tous ces algorithmes ont été conçus et optimisés par rapport aux images LDR. Ces images stockent des valeurs gamma-encodées [0,255] et sont généralement représentées par un nombre entier de 8 bits. Au contraire, les pixels HDR sont représentés par des valeurs réelles proportionnelles à et proportionnels à la luminance physique de la scène, exprimée en cd/m 2 et en et pouvant varier jusqu'à 10^5 cd/m 2 par jour ensoleillé[91]. Par conséquent, les images HDR ont en grande partie des intensités de pixels variables. Il soulève donc naturellement la question de savoir comment ceffectuer une analyse d'images HDR pour les algorithmes HDR l'analyse d'images pour les algorithmes d'extraction de caractéristiques. En d'autres termes simples, il n'est pas clair si les images HDR peuvent être utilisées directement avec de tels algorithmes.

Une alternative serait d'optimiser chaque méthode d'extraction de caractéristiques pour les images HDR. Mais il serait tout à fait impraticable et encombrant, en particulier pour les programmes d'apprentissage existants qui nécessiteraient une grande quantité de données HDR calibrées géométriquement. Sans compter que cela pourrait rendre difficile leur intégration dans des méthodes de vision par ordinateur de niveau intermédiaire et de haut niveau. Dans cette thèse, nous optons donc pour une autre solution. Nous nous concentrons sur les images HDR en entrée et explorons quelle est la meilleure façon d'utiliser de telles images dans des algorithmes d'extraction de caractéristiques optimisés pour le LDR.

Certaines études fondées sur le HDR[2, 11] ont récemment étudié l'impact de l'utilisation d'images HDR sur la performance de détection des caractéristiques. Puisque les algorithmes sont optimisés en LDR, ils faut d'abord convertir le contenu HDR en une image LDR à l'aide de certains opérateurs de Tone Mapping ("TMOs") et ensuite appliquer des techniques de détection de caractéristiques. Ces études, cependant, n'explorent pas d'autres modalité de l'utilisation du HDR (linéaire, par exemple) et n'étudient pas non plus l'impact de l'utilisation de différentes types des TMO existants.

La recherche en imagerie HDR a toujours été abordée d'un point de vue perceptuel. Par conséquent, tous les modes d'utilisation du HDR, appelés "modalités" dans cette thèse, ont été adaptés aux attributs de la vision humaine[13], par exemple la préservation de l'esthétique de l'image, contraste etc. Une façon courante d'évaluer le contenu du HDR est le tone mapping. Par définition, les TMO sont les modèles visant à cartographier le contenu HDR dans une représentation LDR 8 bits appropriée pour l'affichage du contenu sur des écrans des écrans standards. Par exemple, une technique populaire implique la compression de la luminance estimée, par exemple, en utilisant des filtres de préservation des bords tels que les filtres bilatéraux [29] à partir de scènes HDR afin de produire un résultat de tone-mapping visuellement agréable.

Sur le plan conceptuel, les objectifs perceptuels ne sont pas liés aux critères de rendement propres à la vision par ordinateur, comme la note de précision pour la correspondance des caractéristiques. Contrairement à la perception visuelle, les méthode d'extraction de caractéristiques suivent des conceptions strictes pour développer l'invariance dans les informations au niveau des pixels peu localisés, comme l'histogramme des orientations de gradient. Par conséquent, même si les images cartographiées sont des images LDR avec un meilleur contraste sont optimales pour extraire des caractéristiques visuelles robustes.

Par conséquent, il n'est pas clair quelle est la meilleure façon d'utiliser les images HDR ; les images HDR linéaires, une autre forme de HDR codé ou la quantification de l'information ou encore en passant par une représentation LDR, par exemple en utilisant les TMO ?

Dans ce qui suit, nous développons les aspects mentionnés ci-dessus en correspondance avec les chapitres de la thèse.

8.1.2 Chapitre 3

Des conditions d'éclairage défavorables peuvent détériorer considérablement le rendement des détecteurs et des descripteurs de points clés dans l'imagerie LDR conventionnelle. Plusieurs modèles de normalisation locale et globale ont été conçus pour obtenir de meilleures caractéristiques invariantes de luminance. Mais ces techniques sont quelque peu inefficaces dans la pratique. La mauvaise performance de ces algorithmes s'explique principalement par la perte ou le changement dans les configurations spatiales des détails présents dans une scène. L'imagerie HDR permet de dépasser ces limites et, par conséquent, d'améliorer le rendement de l'extracteur de caractéristiques grâce à sa plage dynamique plus large qui permet de capturer des détails dans les régions sombres et claires. Cependant, il n'est pas clair quelles sont les meilleures méthodes pour employer le HDR et ces gains sont significatifs sur un ensemble de données réelles.

Dans le chapitre 3, nous examinons le potentiel du HDR pour les étapes d'extraction des caractéristiques, c'est-à-dire la détection et la description des points clés, et en particulier, nous abordons les questions de recherche suivantes:

1. le HDR est-il capable de réaliser des gains quantitatifs substantiels en termes de stabilité des caractéristiques ? aux changements de luminance par rapport aux LDR ? ces gains sont-ils cohérents ?

2. quelle est la meilleure façon d'utiliser de telles images HDR, c'est-à-dire la luminance réelle directe ou la luminance réelle directe. HDR converti au format LDR par l'intermédiaire d'un opérateur de tone mapping (TMO) afin d'être compatible avec les techniques d'extraction de caractéristiques standard ? Pour répondre à ces questions, un cadre d'évaluation est fourni dans ce chapitre.

Initialement, nous construisons un ensemble de données d'images HDR et LDR, composé de deux configurations, chacune éclairée avec sept et huit conditions d'éclairage différentes, respectivement. L'ensemble de données constitue un défi de taille en termes de réflexion de la texture des objets, de présence d'ombres et de variété d'éclairages sources. Pour chaque scène d'éclairage, nous considérons ensuite un certain nombre de formats de codage d'image, y compris les valeurs HDR linéaires ou codées perceptuellement, la meilleure exposition LDR subjectivement, et plusieurs images locales ou globales cartographiées en tons. Ensuite, nous détectons les caractéristiques de chaque scène d'éclairage et nous calculons la répétabilité standard des points d'intérêt détectés dans tous les autres réglages d'éclairage, afin d'estimer la stabilité moyenne des caractéristiques. Voici accompli en utilisant deux points d'angle populaires (Harris) et des détecteurs de taches (SURF). Certains cadres antérieurs pour l'évaluation des détecteurs et des descripteurs ont été proposés dans la littérature, comme nous l'avons vu à la section 2. Les études sur les receveurs font état d'un augmentation du nombre de points de caractéristiques détectés à l'aide de modalités fondées sur le HDR plutôt que sur le LDR. Cependant, le nombre de

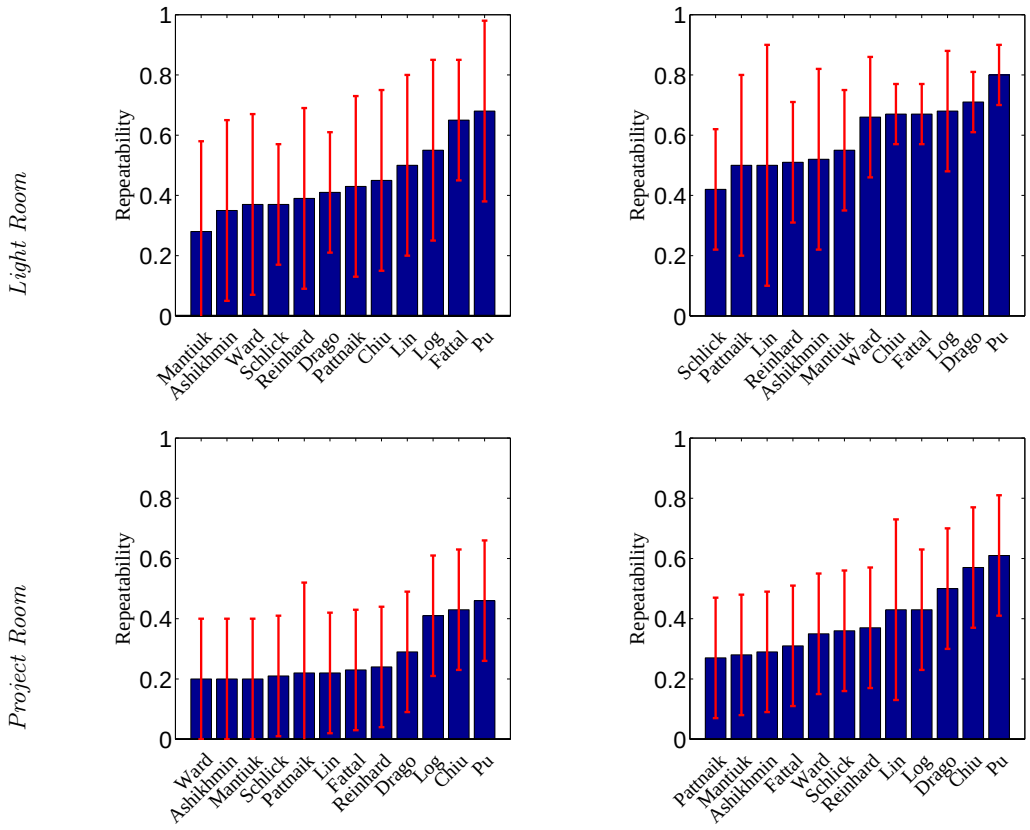


Figure 8.2 – RR moyen enregistré par différents formats sur le LDR.

points de caractéristiques détectés n'est pas en soi un indicateur suffisant de détection la performance. En outre, sur la base de leurs résultats, il est difficile de tirer des conclusions précises sur ce qui fait que certaines modalités de HDR fonctionnent mieux que d'autres.

Le chapitre 3 met l'accent sur les mesures standard de la stabilité des caractéristiques sous éclairage variable et sur l'analyse de la performance de nombreuses approches populaires de cartographie tonale qui ont été évaluées à fond d'un point de vue perceptuel, mais dont l'efficacité dans l'extraction des caractéristiques n'a pas été étudiée jusqu'à présent. De plus, nous comparons explicitement les étapes d'extraction directe de caractéristiques sur les images HDR avec un tonemap-then-extract l'approche.

L'analyse fondée sur des mesures quantitatives du rendement de la détection des points clés et de la détection des points clés. Dans la figure 8.2 et la table 8.1, l'appariement des scores sur différentes scènes confirme le potentiel des techniques HDR sur une seule exposition LDR. Pour la détection et l'appariement, nous observons les valeurs linéaires. Les valeurs HDR sont inappropriées pour être utilisées pour des tâches de reconnaissance visuelle optimisées LDR. Dans le cas des TMO, nous observons que leur performance varie selon le type de scène, en montrant la nature de leur dépendance à l'égard du contenu. De plus, nous constatons que toutes les les TMOs locales produisant des résultats très attractifs ne sont pas nécessairement la meilleure option pour les tâches d'analyse d'images. Plus

Repr.	Feature Extraction Schemes				**Avg/Repr.**
	SIFT	SURF	BRISK	FREAK	
LDR	55	62	60	61	59.5
RNG	69	70	71	65	67.5
DR	72	72	71	**73**	72
RN	72	70	**73**	72	72
MA	74	**75**	62	62	68.3
FA	68	67	62	66	65.8
CH	68	71	64	66	67.3
DU	64	72	68	71	68.8
HDRLog	**75**	66	67	68	69
HDRLin	44	30	50	41	41.5
Avg/Schemes	66.8	65.6	65.5	65	

Table 8.1 – Mean Average Precision (mAP %) scores pour les 10 représentations considérées en utilisant 4 schémas d'extraction de caractéristiques. La moyenne des notes est calculée sur 4 ensembles de données de changement d'éclairage. Le score mAP le plus élevé pour chaque schéma est indiqué en **gras**.

intéressant encore, nous avons également observé que les TMOs locales. avec un taux de répétabilité très élevé pour la détection des caractéristiques ne sont pas nécessairement les meilleurs. lorsque le pipeline d'extraction de toutes les caractéristiques est pris en compte. Pour un test individuel Dans les cas de paires, nous ne trouvons aucune modalité qui est absolument surperformante. Il reste donc il n'est pas clair si les pixels HDR doivent être codés de manière approximativement linéaire par rapport à la perception. ou directement à l'aide des fonctions existantes. Par conséquent, un moyen plus optimal de modalité doit être conçue. (Voir le chapitre 3)

8.1.3 Chapitre 4

Les TMO ont traditionnellement été conçus pour afficher les images HDR d'une manière perceptiblement favorable et surtout pour préserver les attributs de la vision humaine tels que l'esthétique de l'image et le contraste perceptuel. Cependant, lorsque ces images cartographiées par ton doivent être utilisées pour des tâches de vision par ordinateur telles que la détection de points clés, ces approches de conception sont sous-optimales. et doit être recalibré. Il n'existe pas de travaux connexes dans la littérature, qui visent à concevoir une technique de cartographie des tons optimisée pour la détection ou à comprendre les critères qui s'y rattachent.

Dans ce chapitre, nous abordons le problème de la conception optimale des TMO pour la détection des points clés. tâche. Plus précisément, nous examinons les questions suivantes :

1. quels sont les facteurs à prendre en compte dans la conception de la TMO lors du ciblage des tâches de détection des points clés ? et

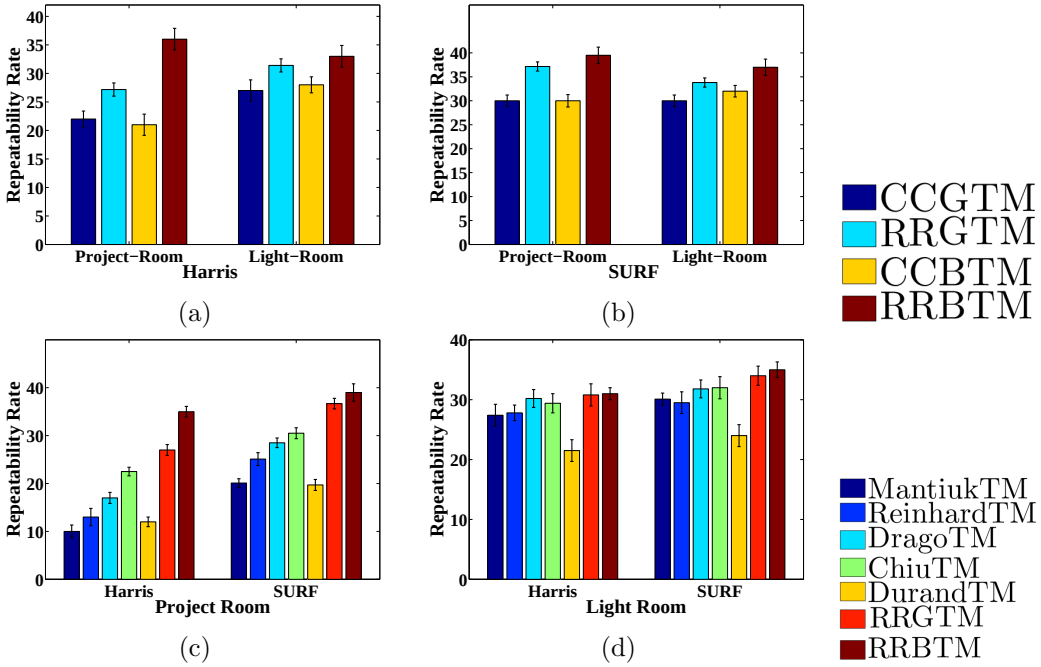


Figure 8.3 – Average RR et standard deviation pour les approches optimisées basées sur la corrélation et la réponse en utilisant respectivement un détecteur Harris et un détecteur SURF. Rangée 2. (c) et (d) Score moyen de répétabilité et écart-type pour les modèles de réflectance (GTM et BTM) et autres TMs couramment utilisés pour Project Room et Light Room dataset

2. comment peut-on nous optimisons un TMO pour de telles tâches sous des variations d'éclairage drastiques.

Pour répondre aux questions susmentionnées, ce chapitre traite dans un premier temps de la sous-optimalité des TMOs existantes et en déduit des lignes directrices pour concevoir un TMO optimisé au niveau des points clés.

A cette fin, premièrement, une comparaison est faite entre l'optimisation des paramètres TMO existants en ce qui concerne : a) le taux de répétabilité RR et b) le coefficient de corrélation CC entre le taux de répétabilité RR et le coefficient de corrélation CC entre des images cartographiées de la même scène avec des variations d'éclairage. CC mesure la statistiquetique entre une paire d'images ton sur ton. L'objectif ici est de déterminer si l'optimisation d'un TMO par rapport à RR conduit à une plus grande stabilité des points clés par rapport à la similarité par pixel (en utilisant CC) entre les images cartographiées par ton. Les résultats sont indiqués dans Figure 8.3.

En nous basant sur les observations de l'étude de l'optimalité, nous présentons dans ce chapitre une nouveau TMO adaptatif basé sur l'apprentissage pour une détection robuste des points clés, nommé DetTMO. La méthodologie globale de DetTMO est illustrée dans Figure 8.4. Le cadre que nous proposons vise à améliorer la détection répétée d'emplacements de points clés clairsemés (p. ex. les coins) dans des zones à contraste élevé de scènes subissant

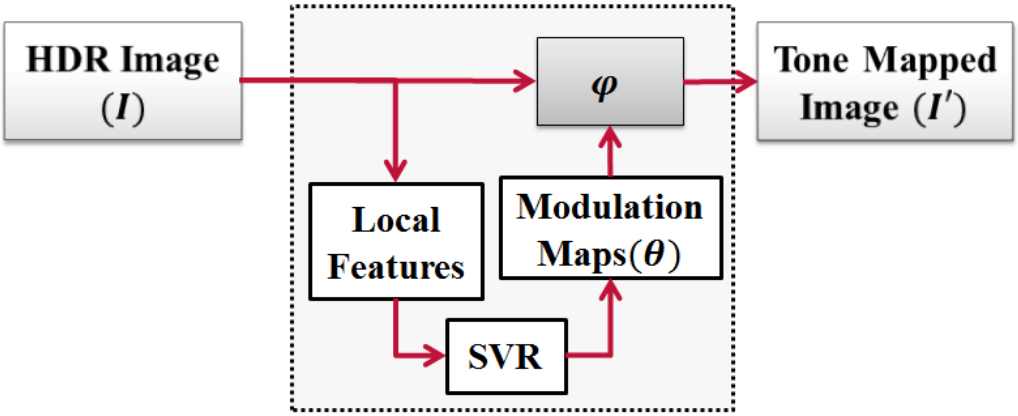


Figure 8.4 – *Learning based DetTMO.*

des transitions complexes d'éclairage du monde réel, comme le changement jour/nuit. Pour ce faire, nous introduisons d'abord un TMO adaptatif qui peut être odulé localement, c'est-à-dire que ses paramètres peuvent varier en pixels.

Ensuite, la modulation par pixel est dérivé au moyen d'un modèle invariant d'illumination appris. Dans ce contexte, nous formons un régisseur de vecteur de support (SVR) pour prédire les cartes de modulation par pixel désirées en en utilisant le contenu HDR linéaire à partir de scènes capturées dans des conditions d'éclairage variables. Les modèles fondés sur l'apprentissage ont rarement été utilisés pour concevoir des modèles optimisés en fonction des points clés. TMOs. Par conséquent, il n'y a pas d'ensemble de données standard pour former ou tester n'importe quel modèle dans le cadre de l ce contexte. Dans ce chapitre, nous surmontons cette difficulté en proposant une détection simple-modèle de maximisation des similitudes pour générer des échantillons de formation appropriés tique entre une paire d'images ton sur ton. L'objectif ici est de déterminer si l'optimisation d'un TMO par rapport à RR conduit à une plus grande stabilité des points clés par rapport à la similarité par pixel (en utilisant CC) entre les images cartographiées par ton.

De plus, nous proposons un ensemble de données HDR de 8 scènes d'images prises à l'intérieur et à l'extérieur avec différentes variations d'éclairage, illustrée dans Figure 8.5

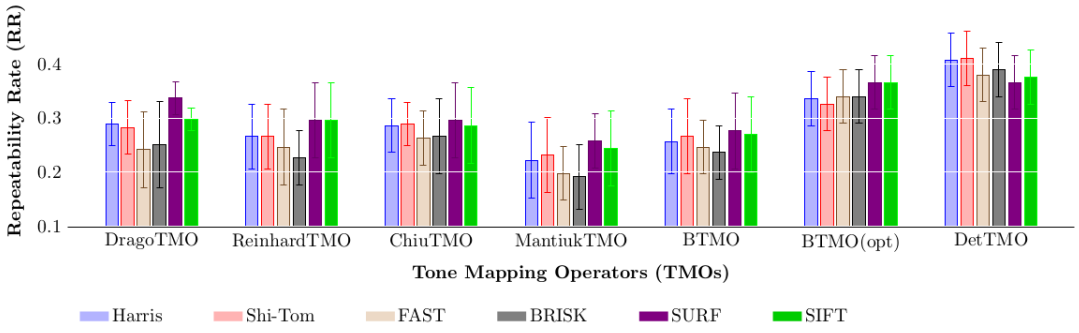


Figure 8.6 – Average Repeatability Rates (AvgRR) calculée sur différents TMOs en utilisant divers schémas de détection de points clés. La moyenne est calculée sur toutes les scènes de test.

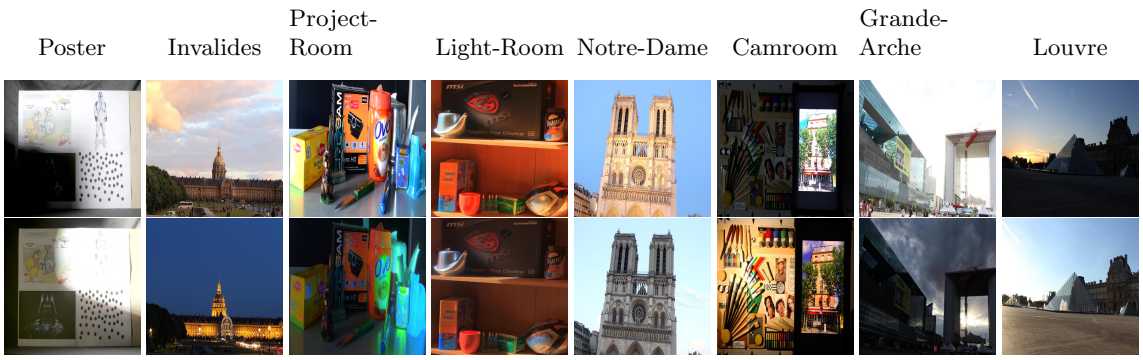


Figure 8.5 – Exemples d'images de *HDR dataset*. *HDR Dataset* est composé de 8 scènes de différents endroits intérieurs et extérieurs.

Nous évaluons la performance du cadre proposé en termes de répétabilité des points clés pour les détecteurs de pointe. Dans Figure 8.6, nous calculons d'abord le RR de toutes les scènes pour chaque TMO considéré et ensuite la moyenne pour calculer le taux de Average Repeatability Rate (AvgRR). Nous observons que pour l'un ou l'autre détecteur (coin ou blob), notre modèle proposé surpasse tous les autres TMOs (basés sur la perception ou sur les points-clés). De plus, les écarts-types plus faibles observés avec notre proposition de TMO montrent une plus grande stabilité des points clés que d'autres TMO basés sur la perception. Bien que notre algorithme ait été optimisé pour les coins, il donne des performances comparables ou meilleures par rapport à d'autres méthodes sur les détecteurs de taches. Cela s'explique en partie par la mise en Åuvre à l'échelle unique des détecteurs à goutte utilisés dans cette évaluation. Cependant, les performances peuvent différer lorsque la détection de blob multi-échelle est prise en compte.

Nos résultats expérimentaux démontrent l'efficacité de notre proposition de TMO

Figure 8.7 – *Repeated Keypoints*. Row I: 2 images HDR de la scène *Invalides* prises à différentes heures du jour. Les images HDR sont affichées après la mise à l'échelle du journal [27]. Row II: les points clés répétés en utilisant notre DetTMO proposé (66 points clés répétés sur les 200 points clés les plus forts). Row III: les points clés répétés à l'aide de Reinhard TMO (7 points clés répétés sur les 200 points clés les plus forts). Row IV: les points clés répétés à l'aide de MantiukTMO (5 points clés répétés sur les 200 points clés les plus forts).

adaptatif basé sur l'apprentissage, qui offre une plus grande stabilité des points clés par rapport aux TMO de pointe basés sur la perception illustrée dans Figure 8.7. (Voir le chapitre 4).

8.1.4 Chapitre 5

Les TMO conventionnels se sont révélés sous-optimaux pour la tâche d'extraction de caractéristiques, qui comprend une étape de détection. Jusqu'à présent, nous avons tiré parti de l'apprentissage des points clé pour concevoir un TMO optimal uniquement pour une détection stable. Dans ce chapitre, nous avons traiter de l'ensemble du pipeline d'extraction d'entités, y compris l'étape de description, afin de concevoir un TMO optimal pour une adaptation efficace de l'image. Plus spécifiquement, l'objectif de ce chapitre est de trouver

un TMO optimal qui peut améliorer l'extraction de caractéristiques stables pour des scènes avec contenant des transitions complexes d'illumination du monde réel, comme le changement jour/nuit. Dans ce but, le chapitre propose d'abord une conception TMO descripteur-optimale, appelée DesTMO, qui vise à uniquement l'extraction de descripteurs invariants (autant que possible) à partir de zones à fort contraste des scènes. Plus tard, nous introduisons un TMO optimal, OpTMO, pour une chaîne complète d'extraction de caractéristiques (comprenant à la fois des détecteurs et des descripteurs) qui améliore simultanément les taux de détection et la correspondance des caractéristiques extraites des scènes HDR. Les deux tâches proposées, à savoir DesTMO et OpTMO, suivent un paradigme basé sur l'apprentissage similaire à celui de DetTMO dans le chapitre 4, mais avec des objectifs de conception entièrement différents.

En résumé, ce chapitre présente,

1. descripteur-optimal DesTMO qui facilite l'extraction des descripteurs invariants de luminance.

2. une OpTMO adaptative locale, optimisée pour l'appariement d'images, qui traite collectivement les étapes de détection et de description des pipelines d'extraction d'entités.

3. une méthode efficace pour générer des échantillons de formation appropriés afin de contourner la difficulté de former les SVR dans le contexte de DesTMO et OpTMO respectivement. En outre, nous proposons leurs fonctions objectives de substitution différentiables.

4. une évaluation de DesTMO et OpTMO par rapport aux méthodologies l'état de l'art. De plus, nous montrons un scénario applicatif de localisation d'objets.

La conception de DesTMO est motivée par le TMO optimal du détecteur du chapitre précédent où des gains significatifs en taux de répétabilité [70] ont été observés lorsque les paramètres TMO optimaux (contrôlant la forme et la taille de TMO) ont été appris en pixels. Cependant, nous nous sommes principalement concentrés sur la conception d'un modèle de cartographie des tons pour les tâches de détection de points-clés en coin. Alors qu'ici, nous nous sommes concentrés sur la conception d'un modèle de cartographie des tons par considérer un problème différent, c'est-à-dire une TMO optimale pour l'extraction de descripteurs discriminatoires.

Pour concevoir DesTMO, on introduit d'abord une fonction de cartographie des tons, qui peut être modulée localement en variant spatialement ses paramètres. La méthodologie globale de DesTMO est illustrée dans Figure 8.8. Ses cartes de paramètres sont prédites au moyen d'un modèle de guidage à illumination invariante. Notre modèle de guidage est piloté par le SVR et s'appuie sur les caractéristiques basées sur l'orientation des gradients qui sont extraits des patchs densément échantillonnés du contenu HDR. Contrairement à la détection

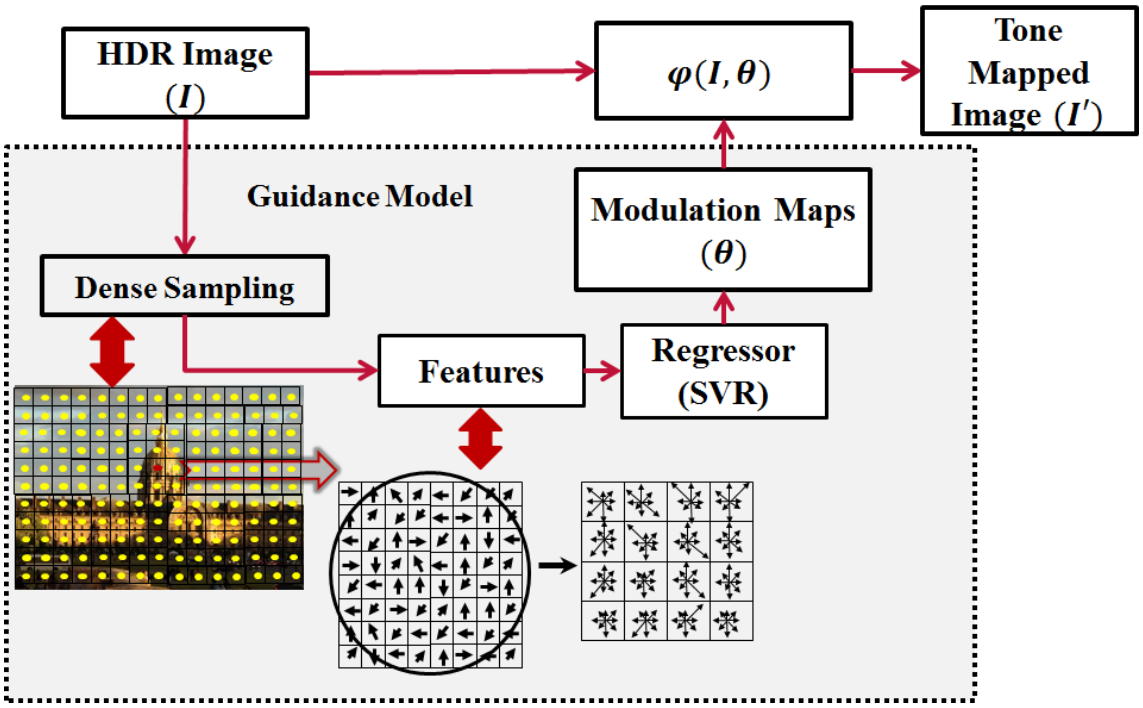


Figure 8.8 – L'architecture de DesTMO.

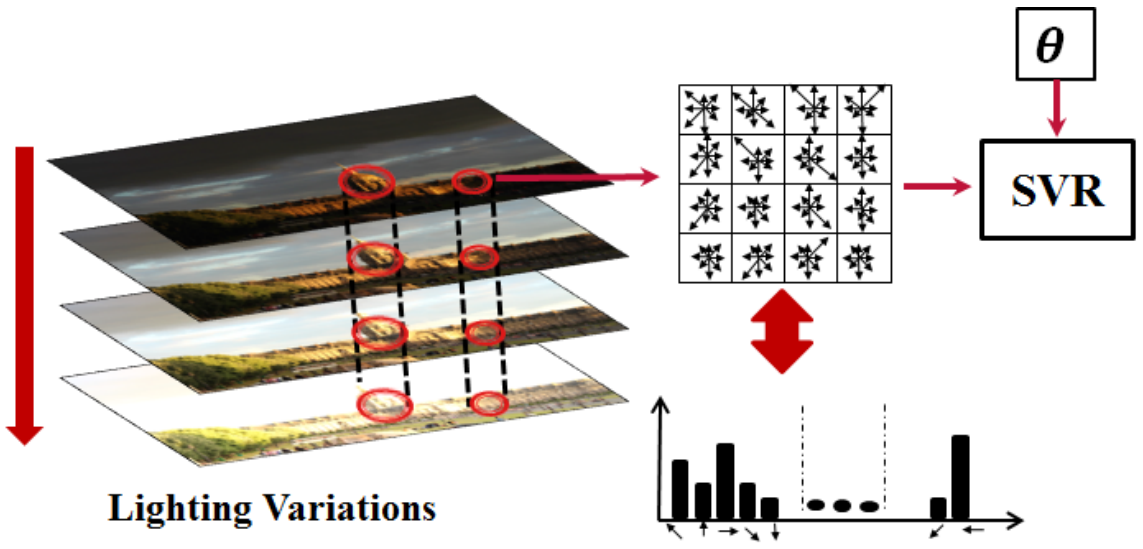


Figure 8.9 – *Apprentissage DesTMO*

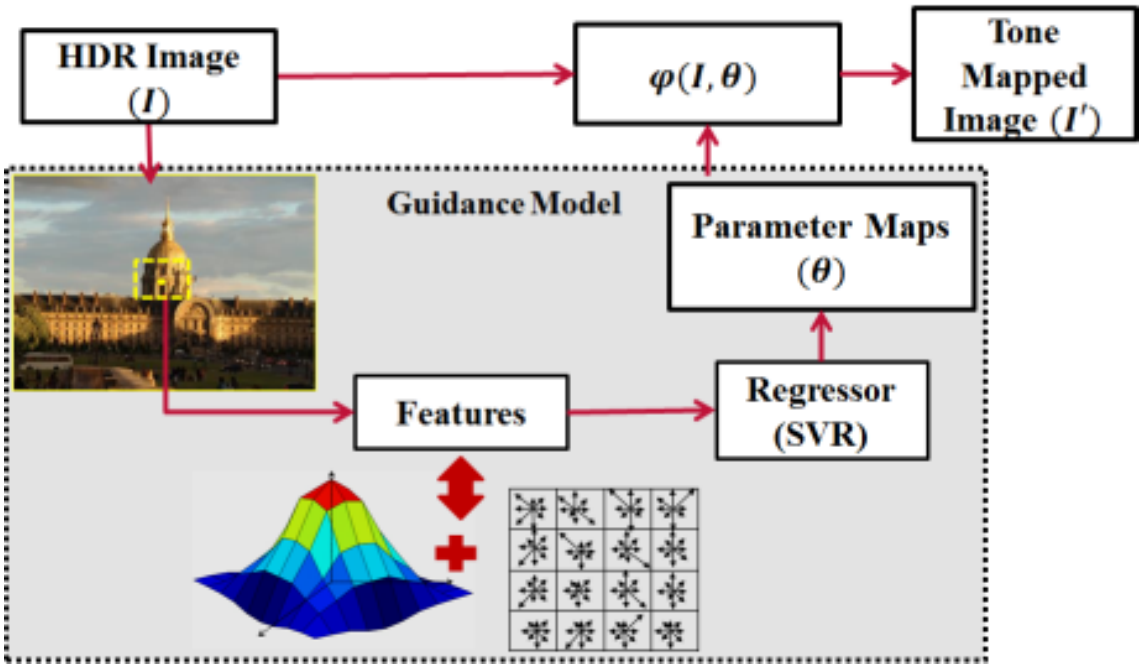


Figure 8.10 – L'architecture de OpTMO.

d'angle, l'extraction de descripteur dépend du grand ensemble des pixels du voisinage (ou patch) qui sont traités ensemble pour formuler la signature unique discriminante. C'est pourquoi nous proposons d'apprendre les paramètres TMO localement, mais en nous basant sur les informations au niveau des patch des scènes. Plus précisément, puisque chaque descripteur est limité à une taille de patch. Comme 16×16 dans SIFT et SURF, nous apprenons les paramètres TMO sur des patchs de même taille.

Puisqu'il n'y a pas d'ensemble de données standard pour former ou tester un modèle pour DesTMO, nous proposons une approche simple de descripteur de maximisation de similarité pour générer des échantillons de formation appropriés. La fonction objective vise à maximiser les similitudes des descripteurs s'ils sont extraits d'images du même endroit mais avec des variations d'éclairage. La filière de formation utilisant l'approche par paires est illustrée dans le Figure 8.9.

DesTMO et DetTMO ne traitent qu'un seul aspect à la fois, à savoir la détection des points-clés ou l'extraction de descripteurs. Ceci n'est pas efficace dans la pratique pour la tâche d'appariement d'images, par *e.g.*, un mauvais détecteur dégrade la correspondance des descripteurs. Notez que l'optimisation d'un TMO en tenant compte simultanément de la détection et de la description des points clés n'est pas trivial, car les objectifs de conception correspondants sont généralement différents et contradictoires.

Par exemple, un TMO optimal pour la détection vise à produire des points de caractéristiques covariantes, alors qu'un TMO optimal pour la description devrait garantir une certaine forme d'invariance des transformations sur un quartier local. De plus, une détection optimale nécessite une localisation précise de la position du point clé, tandis

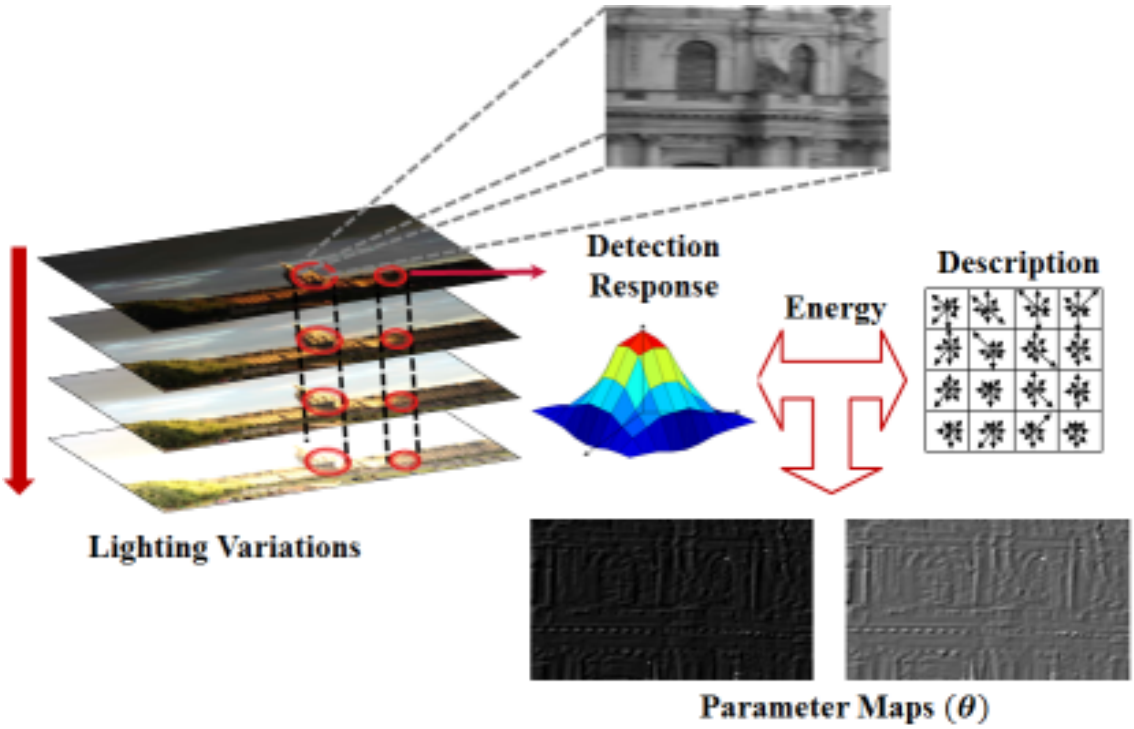


Figure 8.11 – Apprentissage OpTMO.

qu'une description optimale est un processus au niveau du patch. Dans le chapitre 3, nous avons montré que les TMOs qui sont optimales pour la détection ne le sont pas nécessairement lorsque la chaîne complète d'appariement est considérée. Dans ce chapitre, nous présentons un opérateur de cartographie sonore optimale (OpTMO) pour améliorer la détection et l'adaptation des caractéristiques extraites de scènes HDR capturées sous des transitions d'illumination du monde réel complexes. La méthodologie globale de OpTMO est illustrée dans Figure 8.10. Pour OpTMO, nous introduisons d'abord une fonction de tone mapping similaire à DetTMO, qui peut être modulée localement en variant spatialement (en pixels) ses paramètres en fonction des caractéristiques du contenu HDR. Ensuite, nous proposons un modèle d'orientation pour cartographier les caractéristiques locales basées sur le HDR (détection et détection et description) à une faible dimension description à un faible encombrement. La filière de formation utilisant l'approche par paires est illustrée dans le Figure 8.11. Un mauvais détecteur dégrade la correspondance des descripteurs. Notez que l'optimisation d'un TMO en tenant compte simultanément de la détection et de la description des points clés n'est pas trivial, car les objectifs de conception correspondants sont généralement différents et contradictoires.

Enfin, dans Figure 8.12 and 8.13, nous évaluons les deux TMOs proposés sur un ensemble de données HDR de scènes intérieures/extérieures où ils surpassent les TMOs de pointe à travers différents algorithmes d'appariement d'images. Dans Figure 8.14, nous observons que l'optimisation uniquement pour la réponse du détecteur (DetTMO) pourrait

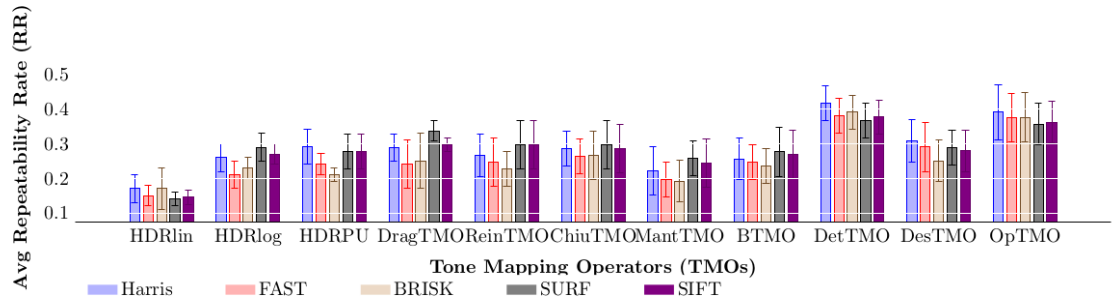


Figure 8.12 – Average RR calculée sur différents TMOs en utilisant divers schémas de détection de points clés. La moyenne est calculée sur toutes les scènes de test.

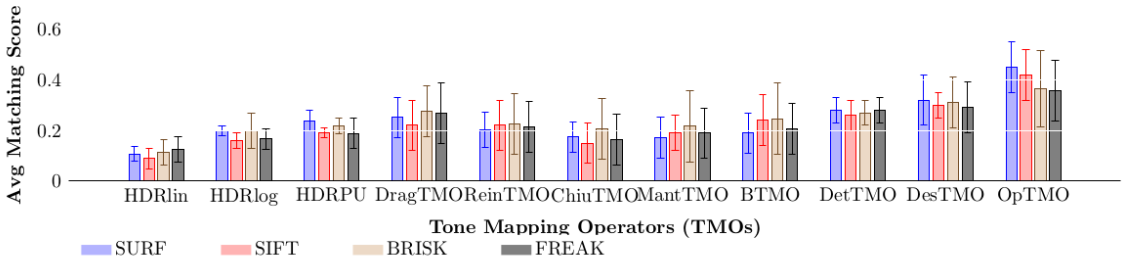


Figure 8.13 – Average Matching Score (MS) calculés sur différents TMOs en utilisant les schémas d'extraction des descripteurs SURF, SIFT, FREAK, BRISK, BRISK. La moyenne est calculée sur toutes les scènes de test.

produire un plus grand nombre de fausses correspondances. D'autre part, l'optimisation en ce qui concerne uniquement l'appariement des descripteurs (DesTMO) ne peut pas assurer une efficacité d'appariement élevée en raison de la faible répétabilité des points-clés. Au lieu de cela, une adaptation efficace de l'image ne peut être assurée qu'en optimisant le TMO par rapport à la chaîne complète d'extraction des caractéristiques, comme dans l'OpTMO proposé.

Nos TMOs optimisés pour les tâches proposées démontrent leur polyvalence lorsqu'ils sont appliqués à différentes approches de détection/description et peuvent donc être directement connectés à diverses applications locales basées sur des fonctionnalités. (Voir le chapitre 5).

8.1.5 Chapitre 6

Avec un objectif donné basé sur les tâches, nous avons jusqu'à présent proposé des modèles en nous appuyant sur des modèles spécifiques caractéristiques d'une fonction de cartographie tonale donnée. En fait, une variante du filtrage bilatéral a été adoptée pour mettre en valeur la capacité d'apprentissage du modèle. Toutefois, tous les TMOs ne sont pas différentiables et, par conséquent, difficiles à apprendre en utilisant les méthodes proposées. De plus, un TMO individuel n'aborde que certaines caractéristiques spécifiques qui pourraient être souhaitées en fonction du contenu. Cela soulève naturellement la question à savoir si pour une cartographie tonale plus générale la fonction peut être formulée de manière à pouvoir être facilement entraînée pour n'importe quelle tâche donnée et à s'adapter pour toutes les scènes du monde réel.

Dans ce chapitre, nous abordons cette question en concevant un TMO générique de bout en bout qui s'adapte à toutes les scènes du monde réel en tenant compte des caractéristiques spécifiques aux tâches souhaitées. Tirant parti d'un large ensemble de données HDR pour des objectifs perceptuels, nous proposons les premières conceptions architecturales DeepTMO (DeepTMO) pour convertir un contenu HDR linéaire en une sortie LDR à haute résolution. Dans la mise en ouvre actuelle nos modèles sont formés pour une tâche perceptuelle, c'est à dire pour donner la sortie la plus réaliste et de haute qualité sans aucun dommage visible à son contenu. Étant l'absence donné qu'une grande quantité de le jeu de données d'images HDR pour la conception d'un TMO basé sur l'apprentissage profond, l'architecture proposée peut également servir de référence pour l'analyse basée sur le HDR. A l'avenir, cela pourrait être exploré en peaufinant le modèle proposé avec un modèle d'apprentissage en cascade spécifique aux tâches en profondeur, par exemple pour l'appariement d'images, la détection des visages, la vidéosurveillance, etc.

Ce chapitre présente 3 réseaux distincts de cartographie tonal basée sur l'apprentissage profond, à savoir DeepTMO-R, DeepTMO-S et DeepTMO-HD. Basé sur des réseaux conditional generative adversarial network (cGAN) [43, 72] comme illustrée dans le Figure 8.15, chacun des modèles proposés prend directement le contenu HDR linéaire et reproduit une

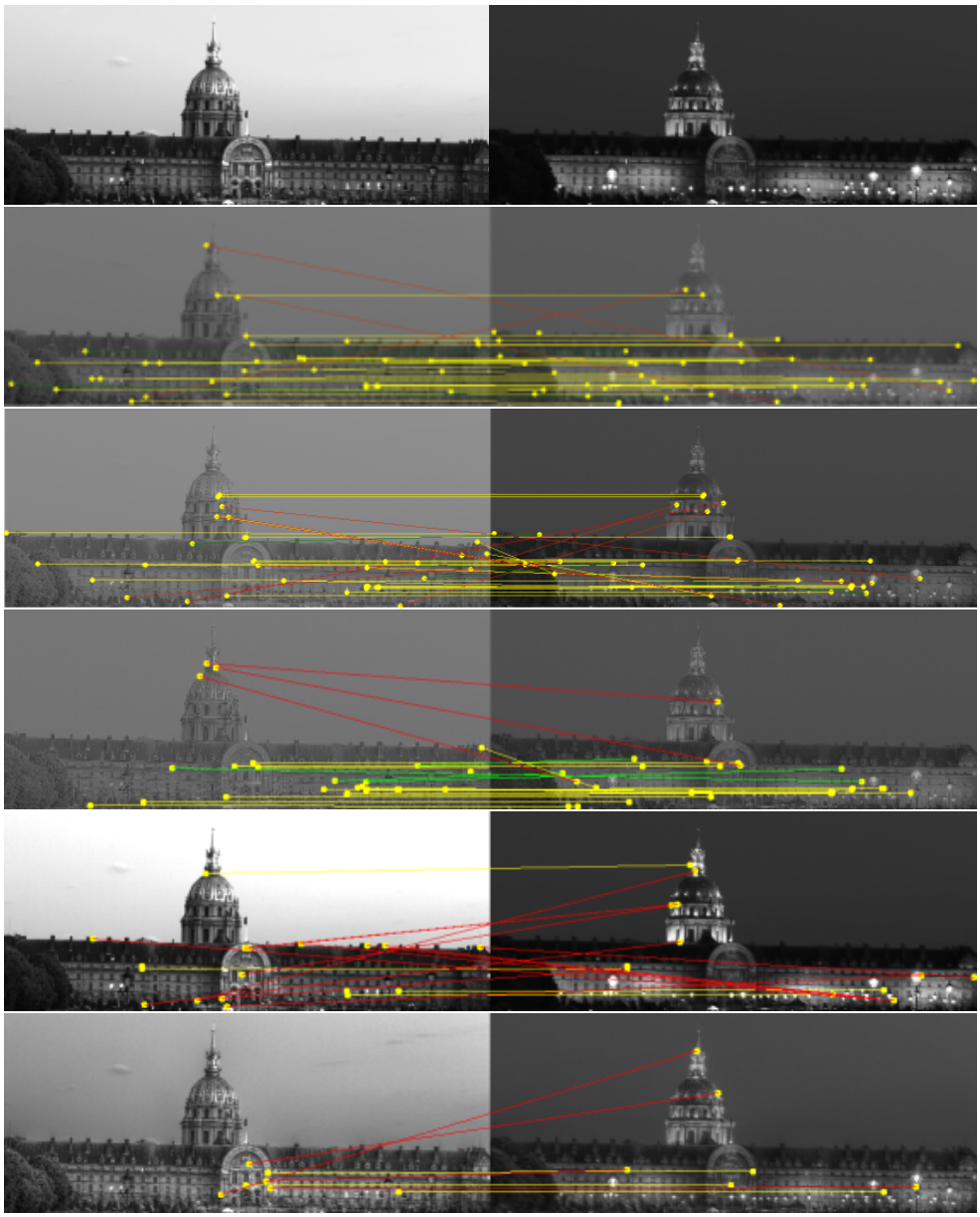


Figure 8.14 – Image Matching II. Correspondance jour/nuit à l'aide de SURF. Row I : 2 images HDR de la scène *Invalides* sont affichées après la mise à l'échelle du journal. Les correspondances correctes et incorrectes sont indiquées par des lignes jaunes et rouges, respectivement. Les lignes vertes représentent le cas particulier d'inadéquation due à une structure répétitive. Row II : l'appariement des caractéristiques à l'aide de notre proposition OpTMO (21 correspondances correctes et 3 incorrectes). Rangée III : en utilisant DetTMO (13 correspondances correctes et 6 incorrectes). Row IV : utiliser DesTMO (11 correspondances correctes et 3 incorrectes). Ligne V en utilisant Reinhard TMO (3 correspondances correctes et 11 incorrectes). Row VI : utiliser MantiukTMO (3 correspondances correctes et 4 incorrectes).

image réaliste visant à imiter le contenu HDR original avec des valeurs de pixels dans la plage[0-255]. Contrairement aux réseaux de convolutional neural network (CNN) explorés dans des travaux antérieurs liés au HDR [30, 32, 41], notre architecture évite l'exigence de

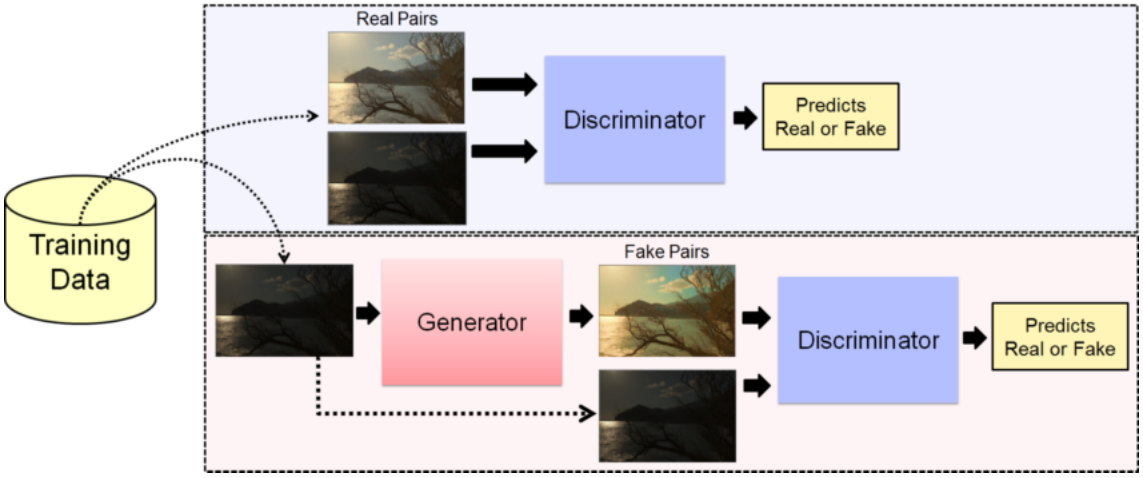


Figure 8.15 – Pipeline de formation de Deep Tone Mapping Operator (Deep TMO) basé sur les GANs. L'ensemble de données apprentisage se compose des HDR d'entrée et de leurs sorties correspondantes les mieux classées en fonction de l'TMQI et de l'indice de qualité du ton. Le discriminateur et le générateur sont formés alternativement, d'abord une étape de regression du discriminateur puis du générateur. Tandis que le discriminateur est formé pour distinguer les paires d'images réelles des fausses, le générateur apprend à tromper le discriminateur en produisant des images en tons synthétiques. Ce faisant, le générateur modélise efficacement la distribution sous-jacente des images réelles de la tonalité de vérité au sol, ce qui donne des résultats de haute qualité une fois l'entraînement terminé.

Figure 8.16 – Nous proposons un TMO basé sur l'apprentissage profond (appelé DeepTMO) qui donne des résultats de haute qualité subjective sur un large éventail d'images HDR à valeur linéaire. Notre variante proposée de cGANs est une architecture multi-échelle qui donne des résultats d'apparence naturelle et sans artefacts en haute résolution. Alors que les TMOs classiques sont sensibles à l'accord des paramètres pour une sortie souhaitée, notre modèle apprend à traiter efficacement une plus large gamme de contenus HDR en modélisant la distribution sous-jacente de toutes les sorties de cartographie des tons cibles disponibles. En concurrence avec les meilleurs résultats des cartes tonales subjectives de qualité supérieure sur 3 types de scènes différentes : claires, nuageuses et sombres, nous montrons surtout la polyvalence de notre méthode qui préserve efficacement les textures, les détails des structures et le contraste. Les résultats détaillés sont présentés aux sections 6 et 7 de chapitre 6. Enfin, notre modèle DeepTMO est assez rapide et prend en moyenne 0,02 seconde pour le tone mapping d'une image HDR de taille 1024×2048.

définir explicitement une fonction de perte spécifique à une tâche. Cela se produit principalement parce que nos réseaux sont formés pour modéliser par eux-mêmes les fonctions de coût adaptées à partir des données de formation sous-jacentes. Nous fournissons les détails architecturaux des trois réseaux dans la Figure 8.17 and 8.18a.

Pour former ces modèles proposés, nous accumulons des données à partir des sources d'images HDR disponibles. Cependant, un défi majeur lors de la formation des modèles découle de l'absence de tout modèle public l'ensemble des données disponibles sur la formation. Sélectionner la vérité terrain par une évaluation subjective plutôt qu'un grand ensemble de données est une tâche très fastidieuse. Elle nécessite donc l'exigence d'un objectif métrique d'évaluation de la qualité qui permet de quantifier les performances de cartographie tonale de chaque TMO pour n'importe quelle scène possible. Pour notre tâche, nous sélectionnons une métrique bien connue, à savoir le Tone Mapped Image Quality Index (TMQI), qui est utilisé pour classer 13 TMOs largement utilisés. En utilisant ceci, pour chaque entrée HDR, nous sélectionnons celui qui se classe le plus haut sur ce score métrique objectif TMQI.

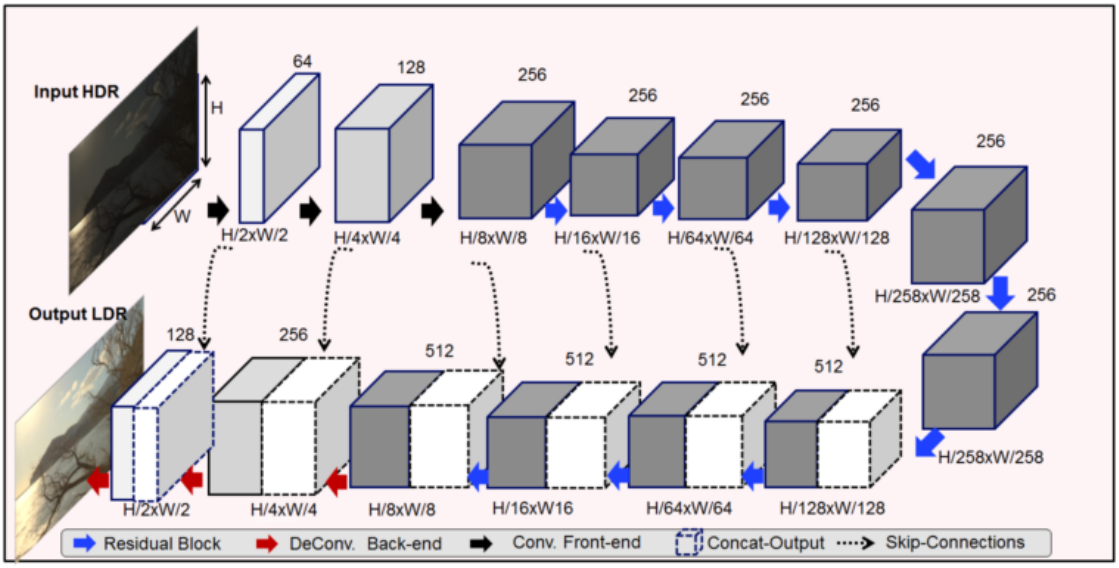
Les résultats visuels produits par nos modèles DeepTMO dans différentes scènes sont illustrés dans le Figure 8.15. Notre DeepTMO apprend implicitement la meilleure caractéristique de tous les TMOs globales, locaux et perceptifs disponibles sur une grande variété de scènes. En un sens, tant par sa conception architecturale que par l'ensemble des données sous-jacentes, il est conditionné pour préserver les données globales (telles que les structures globales, les contrastes et la luminance) ainsi que les détails locaux plus fins (tels que les motifs de texture locale), ce qui permet d'obtenir des résultats de haute qualité visuellement agréables.

Nos modèles de TMO profonds permettent également de surmonter les effets de flou ou de carrelage fréquemment étudiés dans les travaux récents liés au HDR [31, 32], un problème d'intérêt significatif pour plusieurs applications de rendu graphique basé sur l'apprentissage de haute qualité, comme souligné dans[31]. En apprenant simplement une fonction de coût spécifique HDR-to-LDR, les modèles proposés préservent avec succès les caractéristiques de sortie souhaitées telles que le contraste sous-jacent, l'éclairage et les détails minuscules présents dans le HDR d'entrée à l'échelle la plus fine.

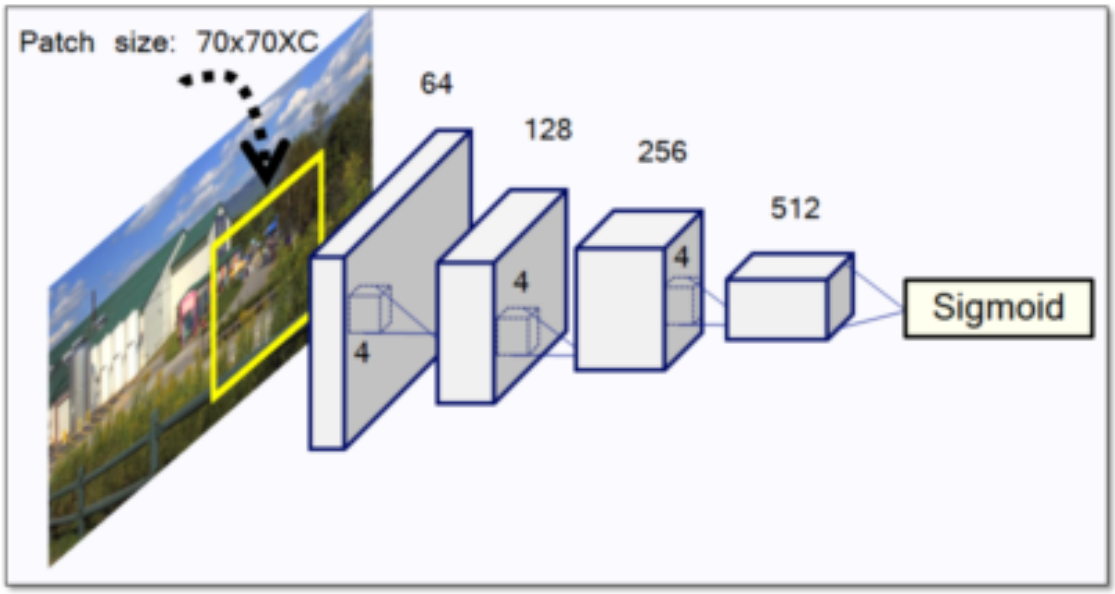
Enfin, nous validons la polyvalence de notre méthodologie par une comparaison quantitative et qualitative détaillée avec les TMOs existantes. Nous comparons les performances quantitatives de DeepTMO-R, DeepTMO-S et DeepTMO-HD avec celles des BestTMO dans le Figure 8.19 et la Table 8.2. Nous démontrons que les TMOs profonds que nous proposons génèrent des images de sortie réalistes de haute qualité et surpassent tous les autres TMOs classiques pour bien généraliser sur un plus large spectre de scènes du monde réel (dans le Figure 8.20).

En un mot,

1. Nous proposons le premier opérateur de tone mapping basée sur l'apprentissage en

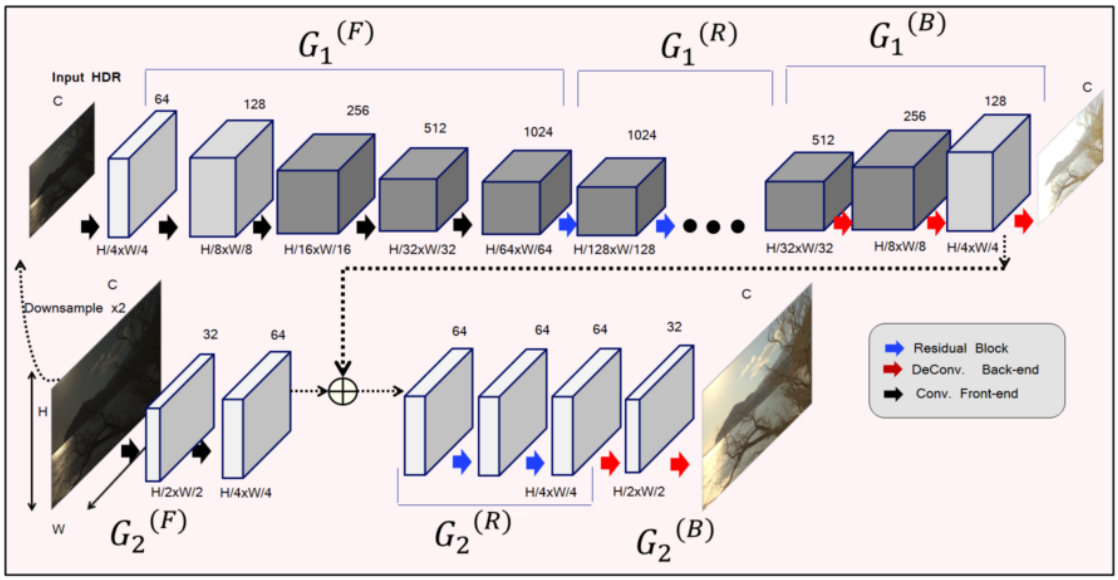


(a) Architecture de générateur avec et sans 'skip-connections'.



(b) Architecture de discriminateur.

Figure 8.17 – Nous présentons l'architecture détaillée du discriminateur et du générateur de DeepTMO-R et DeepTMO-S. La seule différence pour DeepTMO-S est l'ajout de connexions sautées dans le cas d'un générateur. Le générateur est encadré comme une architecture codeur-décodeur, où l'image HDR d'entrée est d'abord transmise à un codeur, qui la sous-échantillonne ensuite en une représentation compacte. Cette représentation est ensuite transmise par le décodeur qui l'échantillonne à la taille du HDR d'entrée. Alors que le codeur se compose du composant frontal Convolution $G^{(F)}$ et des cinq premiers blocs résiduels $G^{(R)}$, le décodeur se compose des quatre blocs résiduels suivants $G^{(R)}$ et du composant déconvolution $G^{(B)}$. La discriminateur se compose d'une architecture patchGAN qui est appliquée à chaque patch sur les paires HDR d'entrée concaténées et les paires LDR de tonalités mappées. La prédiction finale est une moyenne de tous les patchs sur l'image.



(a) L'architecture du générateur pour DeepTMO-HD est une version modifiée de l'architecture du générateur de DeepTMO-R comme le montre la Figure 8.17a. Le générateur DeepTMO-HD est essentiellement une forme de générateur grossier à fin. Tandis que le générateur plus fin G_2 a l'image originale en entrée, l'entrée G_1 est une version échantillonnée $2\times$. Cette image échantillonnée est ensuite effectivement passée à travers les composants suivants $G_1^{(F)}$, $G_1^{(R)}$ et $G_1^{(B)}$ qui sont similaires à ceux du générateur dans DeepTMO-R. La prédiction finale de l'extrémité arrière G_1^B est ensuite concaténée avec la sortie de l'extrémité avant du générateur à plus petite échelle G_2. Ceci est ensuite passé à travers le composant $G_2^{(B)}$ pour produire une sortie avec correspondance des tonalités. Ainsi, notre modèle utilise efficacement l'information à l'échelle plus grossière et à l'échelle plus fine pour faire une prédiction qui permet de mieux retenir la structure globale et les moindres détails des bas niveaux. L'architecture du discriminateur a une architecture identique mais nous lui donnons deux échelles d'entrée différentes, l'original et sa version échantillonnée $2\times$. Cela oblige le générateur à s'occuper à la fois des détails globaux et locaux.

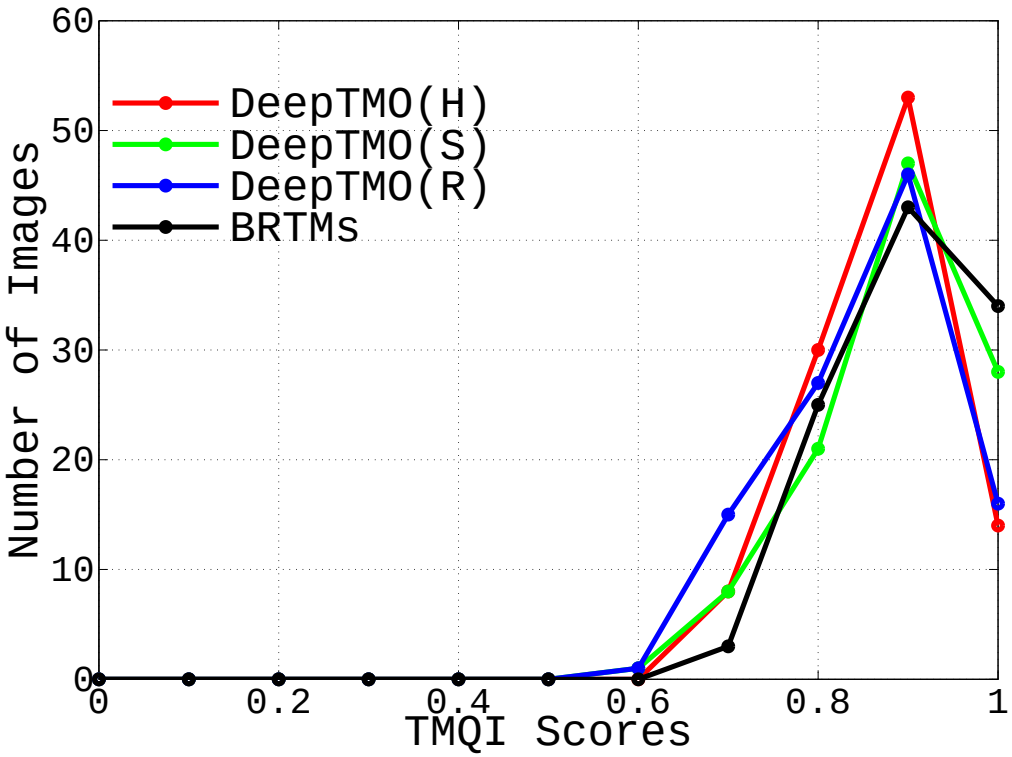


Figure 8.19 – Nous comparons les performances quantitatives de DeepTMO-R, DeepTMO-S et DeepTMO-HD avec celles des BestTMO.

profondeur, qui peut générer des sorties visuellement agréables et réalistes pour une grande variété d'entrées HDR.

2. Nous explorons et comparons 3 architectures cGAN différentes conçues spécifiquement pour générer des sorties LDR à haute résolution pour la cartographie des tonale en préservant les informations structurelles globales ainsi que les détails locaux à grain fin.

3. Nous surmontons le défi de l'indisponibilité des images cartographiées de la tonalité de vérité terrain pour notre ensemble de données HDR en utilisant une métrique objective pour quantifier et classer les divers TMOs.

4. Nous fournissons une comparaison détaillée de la méthodologie que nous proposons avec treize TMOs différents sur un large ensemble de données de 105 images.

5. Notre modèle proposé peut être exploré pour de futures conceptions de TMO optimalaux basé sur l'apprentissage en profond.

(Voir le chapitre 6)

Table 8.2 – *Résultats quantitatifs.* résultats moyens Ã l'TMQI sur un ensemble de 105 images.

TMOs	TMQI
Ward TMO	0.71 ±0.07
Pattnaik TMO	0.78 ±0.04
Log TMO	0.72 ±0.09
Gamma TMO	0.76 ±0.07
Ashikh TMO	0.70 ±0.06
Durand TMO	0.81 ±0.10
Tumblin TMO	0.69 ±0.06
Drago TMO	0.81 ±0.06
Schlick TMO	0.79 ±0.09
Reinh TMO	0.84 ±0.07
Fattal TMO	0.81 ±0.07
Chiu TMO	0.70 ±0.05
Mantiuk TMO	0.84 ±0.06
DeepTMO-HD TMO	0.87 ±0.06
DeepTMO-S TMO	0.88 ±0.07
DeepTMO-R TMO	0.86 ±0.08

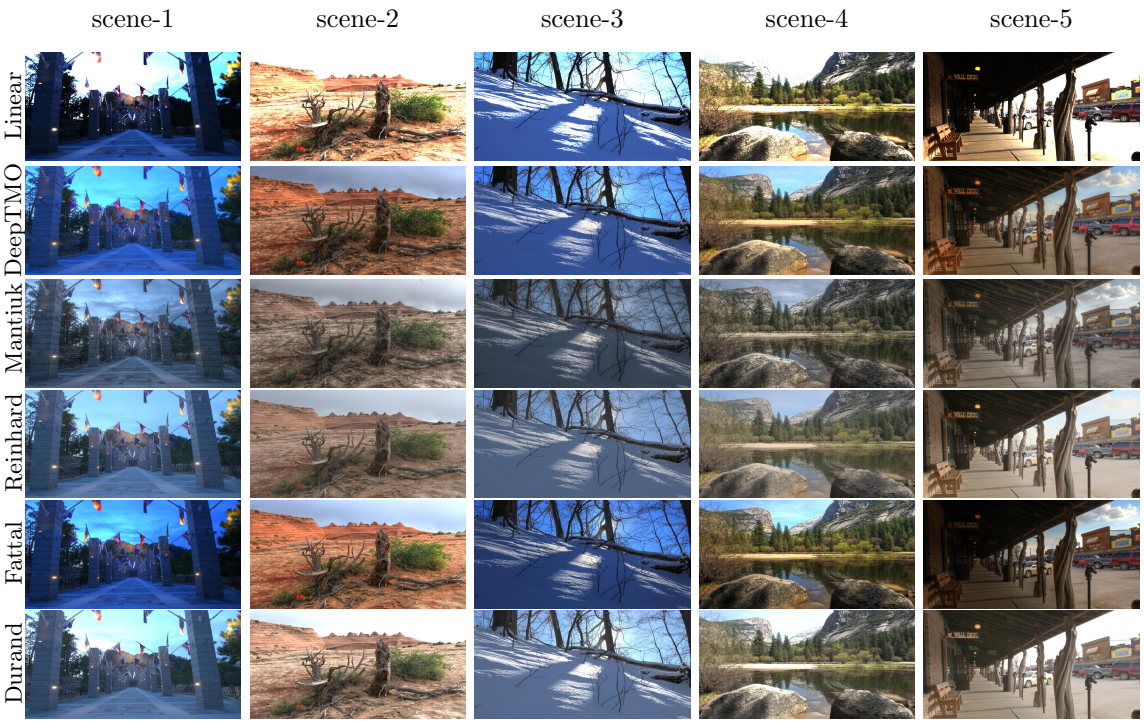


Figure 8.20 – *Resultat Qualitatif de DeepTMO-HD.*

8.1.6 Orientations futures de la recherche

La disponibilité à grande échelle des bases de données d'images et de vidéos du HDR a ouvert la voie à de nouvelles analyses, les perspectives d'avenir de la recherche. Dans cette section, nous discutons de plusieurs extensions possibles de cette thèse.

- Investigation de l'imagerie HDR pour les scènes dynamiques. Comme le montrent les résultats des Figures 4.13 et 5.19, l'information préservée par les images HDR facilite l'extraction de caractéristiques locales très stables et invariantes de luminance. Les scénarios prennent en compte les images HDR qui sont prises à partir de scènes statiques. Cependant, Les scénarios peuvent être dynamiques et, par conséquent, plus difficiles. Cela s'explique principalement par le fait qu'il n'y a pas la même chose. Les combinaisons de transformations physiques telles que les transformations géométriques (rotation, changement de point de vue), des variations déformatrices et dues au bruit des capteurs. Un scénario pratique comprend l'éclairage + l'éclairage des changements de points de vue avec des plates-formes mobiles telles que les drones.

  Les modèles présentés doivent théoriquement s'adapter en fonction de l'invariance des algorithmes d'extraction d'entités, tel qu'illustré à la figure 5.20 pour la rotation planaire. Cependant, pour les problèmes de rotation hors plan qui sont spécifiques à une capture mobile. nos modèles ont besoin d'être recalibrés de manière. Il est à noter qu'il n'y a pas d'équipement de létat les algorithmes d'extraction de caractéristiques locales sont meilleurs sous toutes les transformations [116, 119].

  Par conséquent, au lieu d'apprendre simplement les modèles régresseurs avec un coin de référence et un descripteur basé sur le coin et le descripteur caractéristiques multiples, caractéristiques multiples à partir d'algorithmes d'extraction de caractéristiques, par exemple celles qui ont été évaluées. dans[119], a besoin d'être infusé. De plus, il nécessite un calibrage géométrique approprié qui doit être créé.

- Perception Vs Vision. Au chapitre 6, nous avons proposé les TMOs basés sur l'apprentissage profond pour les TMOs perceptuels. Comme nous l'avons déjà mentionné à la section 6.1, l'une des extensions possibles est la conception d'une tâche optimale. TMO basé sur l'apprentissage profond en affinant le modèle DeepTMO. Cela pourrait nous aider à comparer les résultats obtenus à partir de réseaux profonds tirés de deux objectifs différents (vision perceptuelle et vision par ordinateur) sur un ensemble de données HDR. Puisque la technologie HDR donne une représentation de scènes du monde réel plus proche de l'oeil humain, la comparaison entre les deux modèles ouvrira davantage les possibilités de recherche dans la compréhension plus profonde de ces réseaux. Il motivera en outre la recherche d'explications techniques inspirées par le cerveau humain.

- Analyse HDR dans le domaine temporel. Une grande partie de cette thèse explore

la polyvalence de l'imagerie à gamme dynamique élevée pour améliorer la stabilité des caractéristiques locales dans les images RVB. Cependant, l'une des extensions naturelles est la mise à niveau de l'analyse pour les vidéos HDR. L'information dans le domaine temporel a beaucoup plus de potentiel pour les applications de vision par ordinateur en temps réel telles que les tâches de surveillance. Une grande quantité d'informations perdues dans des scènes à faible contraste peut être reconstruite à l'aide d'une modélisation prédictive en utilisant une dimension temporelle supplémentaire des vidéos HDR. Cela pourrait améliorer les performances dans plusieurs tâches vidéo telles que les applications de surveillance, le suivi en temps réel, l'analyse des gestes et des actions.

- Apprentissage approfondi de l'imagerie HDR avec de petits ensembles de données. En comparaison avec des millions d'images annotées LDR, les ensembles de données HDR accessibles au public sont très petits. Cela limite à son tour l'exportabilité de la technologie HDR. Bien qu'une solution optimale serait de créer un grand ensemble de données HDR avec une évaluation subjective, Si l'on se fie aux vérités du terrain, ce serait une tâche fastidieuse. Une solution alternative peut être l'ingénierie inverse de l'ensemble de données de formation en reconstruisant leur HDR correspondant à l'aide du HDR des modèles récents de TMO inverses basés sur l'apprentissage profond[30, 32]. Une autre alternative pour les travaux futurs peut être de s'appuyer sur la quantité limitée d'échantillons, augmentés de avec des échantillons bruités et ensuite un paradigme d'apprentissage faiblement supervisé[65]. Pour des tâches telles que le mappage HDR vers LDR, un apprentissage totalement non supervisé est également possible, sans donner de paires d'entrées-sorties[120]. L'intuition est de laisser au réseau le soin de décider, par lui-même qui est la meilleure sortie tone-mapping possible simplement en modélisant de manière indépendante la distribution sous-jacente des images HDR d'entrée et des tonalités de sortie.

Abbreviations

BRISK Binary Robust Invariant Scalable Keypoints

CC Correlation Coefficient

cGAN Conditional Generative Adversarial networks

CNNs Convolutional Neural Networks

DeepTMO Deep Learning based Tone mapping operator

DesTMO Descriptor Optimal Tone Mapping Operator

DetTMO Detector Optimal Tone Mapping Operator

FAST Features from Accelerated Segment Test

FREAK Fast Retina Keypoint

Harris Harris Corner Detector

HDR High Dynamic Range

LDR Low Dynamic Range

mAP Mean Average Precision

OpTMO Optimal Tone Mapping Operator for Image Matching

P-R Precision-Recall

RR Repeatability Rate

SIFT Scale-invariant feature transform

SURF Speeded-Up Robust Features

SVR Support Vector Regressor

TMO Tone Mapping Operator

References

[1] W. J. Adams, J. H. Elder, E. W. Graf, J. Leyland, A. J. Lugtigheid, and A. Muryy, "The southampton-york natural scenes (syns) dataset: Statistics of surface attitude," 2016. *Cited in Sec.* 6.6

[2] G. A. Agrafiotis P., Stathopoulou E. and D. A., "HDR imaging for enchancing people detection and tracking in indoor environments," in *In Proceedings of the 10th International Conf. on Computer Vision Theory and Applications (VISIGRAPP)*, 2015. *Cited in Sec.* 1.1, 2

[3] A. O. Akyüz and E. Reinhard, "Color appearance in high-dynamic-range imaging," *Journal of Electronic Imaging*, p. 033001, 2006. *Cited in Sec.* 3.2

[4] M. Ashikhmin, "A tone mapping algorithm for high contrast images," pp. 145–156, 2002. *Cited in Sec.* **??**, 6.6, 6.7, 6.1

[5] T. Aydın, 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, January 2008, pp. 6806–10. *Cited in Sec.* 3.3.2, 3.4.3

[6] M. Azimi, A. Banitalebi-Dehkordi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos, "Evaluating the performance of existing full-reference quality metrics on high dynamic range (hdr) video content," in *International Conference on Multimedia Signal Processing (ICMSP)*, 2014. *Cited in Sec.* 6.6

[7] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers, *Advanced High Dynamic Range Imaging: Theory and Practice*, Natick, MA, USA, 2011. *Cited in Sec.* 2.1.1, 4.3.6, 5.3.1, 5.4.7, 6.5, 6.6, 6.1

[8] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *9th European Conference on Computer Vision (ECCV)*, 2006, pp. 404–417. *Cited in Sec.* 2.2.1, 2.2.2, 3.1, 3.3.1, 3.4.2, 4.2.2, 4.3.6, 5.3.1, 5.4.7, 5.4.8

[9] A. Boschetti, N. Adami, R. Leonardi, and M. Okuda, "An optimal video-surveillance approach for HDR videos tone mapping," in *Signal Processing Conference, 2011 19th European*, Aug 2011, pp. 274–277. *Cited in Sec.* 2.1.2, 2

[10] L. Bottou, *Stochastic Gradient Descent Tricks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 421–436. *Cited in Sec.* 4.3, 4.3.3, 5.2.4, 5.4.2

[11] P. Bronislav, A. Chalmers, and P. Zemcík, "Feature point detection under extreme lighting conditions," in *Spring Conference on Computer Graphics*, 2012, pp. 156–163. *Cited in Sec.* (document), 1.1, 1, 3.1, 3.2, 3.3.3, 3.4.3, 4.1, 4.2.2, 4.3.5, 4.3.6, 5.4.7

[12] 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. *Cited in Sec.* 2.1.2, 3.1, 3.3.1, 3.4.3, 4.2.6, 5.4.3, 5.4.7

[13] M. Cadik, M. Wimmer, L. Neumann, and A. Artusi, "Evaluation of HDR tone mapping methods using essential perceptual attributes," *Computers and Graphics*, pp. 330 – 349, 2008. *Cited in Sec.* 1.1, 2.1.1, 3.3.2, 3.3.3, 3.4.1, 6.6

[14] A. Chalmers and K. Debattista, "Hdr video past, present and future: A perspective," *Signal Processing: Image Communication*, vol. 54, pp. 49 – 55, 2017. *Cited in Sec.* 2

[15] C. Chang and C. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, May 2011. *Cited in Sec.* 4.3.6, 5.3.1, 5.4.6

[16] O. Chapelle and M. Wu, "Gradient descent optimization of smoothed information retrieval metrics," *Information Retrieval*, vol. 13, no. 3, pp. 216–235, 2010. [Online]. Available: http://dx.doi.org/10.1007/s10791-009-9110-3 *Cited in Sec.* 5.4.2

[17] J. Chen, A. Adams, N. Wadhwa, and S. W. Hasinoff, "Bilateral guided upsampling," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, p. 203, 2016. *Cited in Sec.* 6.2

[18] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 1520–1529. [Online]. Available: https://doi.org/10.1109/ICCV.2017.168 *Cited in Sec.* 6.2.1

[19] L. Chermak and N. Aouf, "Enhanced feature detection and matching under extreme illumination conditions with a hdr imaging sensor," in *IEEE 11th International Conference on Cybernetic Intelligent Systems*, Aug 2012, pp. 64–69. *Cited in Sec.* 1, 3.1, 3.3.3, 5.4.7

[20] L. Chermak, N. Aouf, and M. Richardson, "HDR imaging for feature tracking in challenging visibility scenes," *Kybernetes*, pp. 1129–1149, 2014. *Cited in Sec.* 1, 3.1, 3.3.1

[21] 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. *Cited in Sec.* 2.1.1, **??**, 3.3.3, **??**, 4.2, 4.2.1, 4.2.6, 4.3.1, 4.3.2, 4.3.6, 5.2.2, 5.3.1, 5.4.6, 6.6, 6.7, 6.1

[22] Y. Cui, A. Pagani, and D. Stricker, "Robust point matching in hdri through estimation of illumination distribution," in *Pattern Recognition*. Springer Berlin Heidelberg, 2011, pp. 226–235. *Cited in Sec.* 1

[23] M. Database. (2004) Mpi hdr image database. [Online]. Available: http://resources.mpi-inf.mpg.de/hdr/gallery.html *Cited in Sec.* 6.6

[24] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH, New York, USA, 1997, pp. 369–378. *Cited in Sec.* (document), 2, 2.1, 4.3.5, 6.6

[25] J. Dong, N. Karianakis, D. Davis, J. Hernandez, J. Balzer, and S. Soatto, "Multi-view Feature Engineering and Learning," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015. *Cited in Sec.* 5.2.4, 5.4.2

[26] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 658–666. *Cited in Sec.* 6.3.4

[27] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive logarithmic mapping for displaying high contrast scenes," *Computer Graphics Forum*, pp. 419–426, 2003. *Cited in Sec.* (document), 2.1.1, **??**, 3.3.3, **??**, 4.2.6, 4.3.1, 4.3.6, 4.13, 5.3.1, 5.7, 5.4.6, 5.19, 6.6, 6.7, 6.1

[28] F. Dufaux, P. Le Callet, R. Mantiuk, and M. Mrak, *High Dynamic Range Video: From Acquisition, to Display and Applications*. Academic Press, 2016. *Cited in Sec.* 1

[29] 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. *Cited in Sec.* 1.1, 2.1.1, **??**, 4.2, 4.2.6, 5.2.2, 5.4.6, 6.6, 6.7, 6.7.2, **??**, 6.1

[30] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "Hdr image reconstruction from a single exposure using deep cnns," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, p. 178, 2017. *Cited in Sec.* 6.1, 6.2, 6.3.2, 7.2

[31] G. Eilertsen, R. Wanat, R. K. Mantiuk, and J. Unger, "Evaluation of Tone Mapping Operators for HDR-Video," *Computer Graphics Forum*, 2013. *Cited in Sec.* 6.8

[32] Y. Endo, Y. Kanamori, and J. Mitani, "Deep reverse tone mapping," *ACM Trans. Graph.*, Nov. 2017. *Cited in Sec.* 6.1, 6.2, 6.3.2, 6.8, 7.2

[33] ETHyma. (2015) Ethyma database for high dynamic range images. [Online]. Available: http://ivc.univ-nantes.fr/en/databases/ETHyma/ *Cited in Sec.* 6.6

[34] M. Fairchild. (2007) The hdr photographic survey. [Online]. Available: http://www.rit-mcsl.org/fairchild/HDR.html *Cited in Sec.* 6.6, 6.7

[35] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 249–256, Jul. 2002. *Cited in Sec.* 2.1.1, **??**, 3.3.3, **??**, 6.3.4, 6.6, 6.7.2, **??**, 6.1

[36] W. Förstner, T. Dickscheid, and F. Schindler, "Detecting interpretable and accurate scale-invariant keypoints," in *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, 2009, pp. 2256–2263. *Cited in Sec.* 2.2.1, 5.4.2

[37] 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," 2014. [Online]. Available: http://spiedigitallibrary.org *Cited in Sec.* 6.6

[38] H. Gao, H. Yuan, Z. Wang, and S. Ji, "Pixel deconvolutional networks," *arXiv preprint arXiv:1705.06820*, 2017. *Cited in Sec.* (document), 6.7.1, 6.5

[39] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423. *Cited in Sec.* 6.3.4

[40] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand, "Deep joint demosaicking and denoising," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, p. 191, 2016. *Cited in Sec.* 6.2

[41] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 118, 2017. *Cited in Sec.* 6.1, 6.2

[42] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006. *Cited in Sec.* 4.2.1

[43] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680. *Cited in Sec.* 6.1, 6.2.1, 6.5

[44] R. Gupta and A. Mittal, *SMD: A Locally Stable Monotonic Change Invariant Feature Descriptor*. Springer Berlin Heidelberg, 2008. *Cited in Sec.* 1, 1.1, 3.1

[45] C. Harris and M. Stephens, "A combined corner and edge detector," in *In Proc. of Fourth Alvey Vision Conference*, 1988, pp. 147–151. *Cited in Sec.* 2.2, 2.2.1, 3.1, 3.3.1, 4.2.2, 4.2.6, 4.3.3, 4.3.6, 5.4.2, 5.4.7

[46] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004. *Cited in Sec.* 5.4.10

[47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *Cited in Sec.* 6.5

[48] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017. *Cited in Sec.* (document), 6.2.1, 6.2, 6.3.1

[49] T. Jinno, S. Kuriyama, and M. Okuda, "Tone-mapping for an hdr surveillance system using SIFT features," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, Sept 2013, pp. 1–5. *Cited in Sec.* 2.1.2

[50] D. J. Jobson, Z. Rahman, and G. A. Woodell, "Properties and performance of a center/surround retinex," *Image Processing, IEEE Transactions on*, Mar 1997. *Cited in Sec.* 4.2.5

[51] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711. *Cited in Sec.* 6.2.1, 6.3.1, 6.3.4

[52] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2017)*, vol. 36, no. 4, 2017. *Cited in Sec.* 6.2

[53] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, Jun. 2010. *Cited in Sec.* 4.3

[54] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. *Cited in Sec.* 6.5

[55] G. Kontogianni, E. K. Stathopoulou, A. Georgopoulos, and A. Doulamis, "HDR imaging for feature detection on detailed architectural scenes," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 325–330, 2015. *Cited in Sec.* 2.1.2, 3, 3.1

[56] P. Korshunov, M. V. Bernardo, A. M. G. Pinheiro, and T. Ebrahimi, "Impact of tone-mapping algorithms on subjective and objective face recognition in HDR images," in *International ACM Workshop on Crowdsourcing for Multimedia (CrowdMM)*, 2015. *Cited in Sec.* 4

[57] G. Krawczyk. (2006) Mpi hdr video database. [Online]. Available: http://resources.mpi-inf.mpg.de/hdr/video/ *Cited in Sec.* 6.6

[58] G. W. Larson, H. Rushmeier, and C. Piatko, "A visibility matching tone reproduction operator for high dynamic range scenes," *IEEE Transactions on Visualization and Computer Graphics*, pp. 291–306, Oct. 1997. *Cited in Sec.* 2.1.1, **??**, 6.6, 6.1

[59] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, "Evaluation of tone mapping operators using a high dynamic range display," *ACM Transactions on Graphics*, pp. 640–648, 2005. *Cited in Sec.* 2.1.1, 6.6

[60] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint arXiv:1609.04802*, 2016. *Cited in Sec.* (document), 6.2.1, 6.3, 6.2, 6.3.1

[61] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proceedings of the 2011 International Conference on Computer Vision*, ser. ICCV '11, Washington, DC, USA, 2011, pp. 2548–2555. *Cited in Sec.* 2.2.1, 2.2.2, 3.4.2, 4.3.6, 5.3.1, 5.4.7, 5.4.8

[62] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 702–716. *Cited in Sec.* (document), 6.3, 6.2, 6.3.1

[63] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *arXiv preprint arXiv:1703.00848*, 2017. *Cited in Sec.* 6.2.1

[64] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. *Cited in Sec.* 2.2, 2.2.1, 2.2.2, 3.4.2, 4.2.6, 4.3.6, 5.2.4, 5.3.1, 5.4.2, 5.4.3, 5.4.7, 5.4.8

[65] Z. Lu, Z. Fu, T. Xiang, P. Han, L. Wang, and X. Gao, "Learning from weak and noisy labels for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 3, pp. 486–500, March 2017. *Cited in Sec.* 7.2

[66] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models." *Cited in Sec.* 6.4.2

[67] 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, Jul. 2006. *Cited in Sec.* 2.1.1, **??**, **??**, 4.2.6, 4.3.1, 4.3.6, 4.3.7, 5.3.1, 5.3.2, 5.4.6, 6.6, 6.7, 6.7.2, 6.1

[68] R. K. Mantiuk, K. Myszkowski, and H.-P. Seidel, *High Dynamic Range Imaging. Cited in Sec.* 1, 2

[69] H. P. A. M. G. P. Manuela Pereira, Juan-Carlos Moreno, "Automatic face recognition in hdr imaging," pp. 9138 – 9138 – 10, 2014. *Cited in Sec.* 4

[70] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005. *Cited in Sec.* 2.2, 2.2.1, 2.2.2, 2.2.3, 3.4.3, 4.2.1, 5.2, 5.4, 5.4.5, 5.4.7, 5.4.8, 5.4.9

[71] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A Comparison of Affine Region Detectors," *International Journal of Computer Vision*, 2005. *Cited in Sec.* 2.2.3

[72] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014. *Cited in Sec.* 6.1, 6.2.1, 6.3, 6.3.4

[73] T. Mitsunaga and S. Nayar, "Radiometric self calibration," in *IEEE Conference on Computer Vision and Pattern Recognition, 1999.*, 1999, pp. –380 Vol. 1. *Cited in Sec.* 2.1

[74] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814. *Cited in Sec.* 6.4.1

[75] K. S. Ni and T. Q. Nguyen, "Image superresolution using support vector regression," *IEEE Transaction on Image Processing*, vol. 16, no. 6, Jun. 2007. *Cited in Sec.* 4.3

[76] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016. *Cited in Sec.* (document), 6.7.1, 6.5

[77] R. Ortiz, "Freak: Fast retina keypoint," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, ser. CVPR '12, Washington, DC, USA, 2012, pp. 510–517. *Cited in Sec.* 2.2.2, 3.4.2, 5.3.1, 5.4.8

[78] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017. *Cited in Sec.* 6.5

[79] S. Pattanaik and H. Yee, "Adaptive gain control for high dynamic range image display," in *Proceedings of the 18th Spring Conference on Computer Graphics*, ser. SCCG '02. ACM, 2002, pp. 83–87. *Cited in Sec.* 2.1.1, **??**, 6.6, 6.7, 6.1

[80] Pfstools. (2007) Pfstools image database. [Online]. Available: http://pfstools.sourceforge.net/hdr_gallery.html *Cited in Sec.* 6.6

[81] A. A. Rad, L. Meylan, P. Vandewalle, and S. Süsstrunk, "Multidimensional image enhancement from a set of unregistered and differently exposed images," in *Computational Imaging V, San Jose, CA, USA, January 29-31, 2007*, 2007. [Online]. Available: http://lcavwww.epfl.ch/alumni/meylan/ *Cited in Sec.* 6.6

[82] A. Rana, G. Valenzise, and F. Dufaux, "Evaluation of feature detection in HDR based imaging under changes in illumination conditions," in *IEEE International Symposium on Multimedia (ISM), Miami,USA, December, 2015*, 2015, pp. 289–294. *Cited in Sec.* 3.1, 3.4.3, 4.1, 4.2.2, 4.2.4, 4.2.6, 4.3.5, 4.3.6, 5.4.7

[83] A. Rana, G. Valenzise, and F. Dufaux, "An evaluation of HDR image matching under extreme illumination changes," in *The International Conference on Visual Communications and Image Processing (VCIP)*, Chengdu, China, Nov. 2016. *Cited in Sec.* 4.1, 5.4, 5.4.7

[84] A. Rana, G. Valenzise, and F. Dufaux, "Optimizing Tone Mapping Operators for Keypoint Detection under Illumination Changes," in *2016 IEEE Workshop on Multimedia Signal Processing (MMSP 2016)*, Montréal, Canada, Sep. 2016. [Online]. Available: https://hal-institut-mines-telecom.archives-ouvertes.fr/hal-01349708 *Cited in Sec.* (document), 4.3.1, 4.3.2, 5.2.2, 5.3.1, 5.3.2, 5.4.2, 5.4.7, 5.12, 5.16

[85] ——, "Learning-based Adaptive Tone Mapping for Keypoint Detection," in *IEEE International Conference on Multimedia & Expo (ICME'2017)*, Hong Kong, China, Jul. 2017. *Cited in Sec.* (document), 5.4.3, 5.4.6, 5.4.7, 5.12, 5.16

[86] ——, "Learning-Based Tone Mapping Operator for Image Matching," in *IEEE International Conference on Image Processing (ICIP'2017)*. Beijing, China: IEEE, 2017. *Cited in Sec.* (document), 5.4.2, 5.4.6, 5.4.7, 5.16

[87] A. Rana, J. Zepeda, and P. Pérez, "Feature learning for the image retrieval task," in *Computer Vision - FSLCV, Asian Conference on Computer Vision (ACCV) 2014 - Singapore, November 1-2, 2014*, 2014, pp. 152–165. *Cited in Sec.* 4.3.3, 5.2.4, 5.4.2

[88] A. Rana, G. Valenzise, and F. Dufaux, "Evaluation of feature detection in HDR based imaging under changes in illumination conditions," in *IEEE International Symposium on Multimedia, ISM 2015, Miami,USA, December, 2015*, 2015, pp. 289–294. *Cited in Sec.* 6.6

[89] E. Reinhard, "Parameter estimation for photographic tone reproduction," *J. Graphics, GPU, & Game Tools*, vol. 7, no. 1, pp. 45–51, 2002. *Cited in Sec.* 2.1.1

[90] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Transactions on Graphics*, pp. 267–276, Jul. 2002. *Cited in Sec.* 2.1.1, **??**, **??**, **??**, 4.2.6, 4.3.1, 4.3.2, 4.3.6, 4.3.7, 5.2.2, 5.3.1, 5.3.2, 5.4.6, 6.3.4, 6.6, 6.7, 6.7.2, 6.1

[91] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting (The Morgan Kaufmann Series in Computer Graphics)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. *Cited in Sec.* 1.1, 2.1.1

[92] M. Rerabek, L. Yuan, L. Krasula, P. Korshunov, K. Fliegel, and T. Ebrahimi, "Evaluation of privacy in high dynamic range video sequences," in *Applications of Digital Image Processing XXXVII*, Sep. 2014. *Cited in Sec.* 5

[93] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241. *Cited in Sec.* 6.3.2

[94] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proceedings of the 9th European Conference on Computer Vision - Volume Part I*, ser. ECCV'06. Berlin, Heidelberg: Springer-Verlag, pp. 430–443. *Cited in Sec.* 2.2.1, 4.3.6, 5.4.7

[95] C. Schlick, "An adaptive sampling technique for multidimensional integration by ray-tracing," in *Photorealistic Rendering in Computer Graphics*. Springer, 1994, pp. 21–29. *Cited in Sec.* 2.1.1, **??**, 6.6, 6.7, 6.1

[96] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *International Journal of Computer Vision*, pp. 151–172, Jun. 2000. *Cited in Sec.* 1.1, 2.2, 2.2.1, 3.2, 3.3.3, 4.2.1, 4.2.2, 4.2.3, 5.4.2, 5.4.7

[97] S. Shan, W. Gao, B. Cao, and D. Zhao, "Illumination normalization for robust face recognition against varying lighting conditions," in *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, ser. AMFG '03. Washington, DC, USA: IEEE Computer Society, 2003. *Cited in Sec.* 1.1

[98] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. *Cited in Sec.* 5.4.2, 6.3.4

[99] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, Aug. 2004. *Cited in Sec.* 4.3.1, 5.2.3, 5.4

[100] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014. *Cited in Sec.* 6.4.1, 6.5

[101] R. Suma, G. Stavropoulou, E. Stathopoulou, L. V. Gool, A. Georgopoulos, and A. Chalmers, "Evaluation of the effectiveness of HDR tone-mapping operators for photogrammetric applications," *Virtual Archaeology Review*, vol. 7, no. 15, 2016. *Cited in Sec.* 2.1.2, 3, 4.3.6, 5.3.1

[102] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 349–366, 2007. *Cited in Sec.* 4.3

[103] F. Tang, S. H. Lim, N. L. Chang, and H. Tao, "A novel feature descriptor invariant to complex brightness changes," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, 2009, pp. 2631–2638. *Cited in Sec.* 1.1, 3.1

[104] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the Sixth International Conference on Computer Vision*, ser. ICCV '98, Washington, DC, USA, 1998. *Cited in Sec.* 2.2.1, 4.2.1, 4.3.2, 4.3.6, 5.2.2, 5.4.2

[105] J. Tumblin, J. K. Hodgins, and B. K. Guenter, "Two methods for display of high contrast images," *ACM Transactions on Graphics*, pp. 56–94, Jan. 1999. *Cited in Sec.* 2.1.1, 6.6, 6.1

[106] H.-Y. F. Tung, A. Harley, W. Seto, and K. Fragkiadaki, "Adversarial inversion: Inverse graphics with adversarial priors," *arXiv preprint arXiv:1705.11166*, 2017. *Cited in Sec.* 6.2.1

[107] T. Tuytelaars and K. Mikolajczyk, *Local Invariant Feature Detectors: A Survey.* Hanover, MA, USA: Now Publishers Inc., 2008. *Cited in Sec.* 2.2.1, 5.4.7

[108] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016. *Cited in Sec.* 6.4.1, 6.5

[109] A. Vedaldi and B. Fulkerson, "VLFeat: An Open and Portable Library of Computer Vision Algorithms," 2008. *Cited in Sec.* 5.2.4, 5.3.1, 5.4.2

[110] Y. Verdie, K. M. Yi, P. Fua, and V. Lepetit, "TILDE: A temporally invariant learned detector." in *CVPR.* IEEE Computer Society, 2015, pp. 5279–5288. *Cited in Sec.* 1, 1.1, 2.2.1, 3.1, 5.4.2

[111] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," *arXiv preprint arXiv:1711.11585*, 2017. *Cited in Sec.* 6.2.1, 6.3.3

[112] Z. Wang, B. Fan, and F. Wu, "Local intensity order pattern for feature description," in *2011 International Conference on Computer Vision*, Nov 2011. *Cited in Sec.* 1.1

[113] F. Xiao, J. M. DiCarlo, P. B. Catrysse, and B. A. Wandell, "High dynamic range imaging of natural scenes," in *In Tenth Color Imaging Conference: Color Science, Systems, and Applications*, 2002. *Cited in Sec.* 6.6

[114] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 657–667, Feb 2013. *Cited in Sec.* 6.6

[115] R. Yeh, C. Chen, T. Y. Lim, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with perceptual and contextual losses," *arXiv preprint arXiv:1607.07539*, 2016. *Cited in Sec.* 6.2.1

[116] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned Invariant Feature Transform," in *Proceedings of the European Conference on Computer Vision*, 2016. *Cited in Sec.* 1.1, 2.2, 5.4.2, 7.2

[117] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. *Cited in Sec.* 1

[118] B. Zhang and A. J. P., "Adaptive bilateral filter for sharpness enhancement and noise removal," *IEEE Transactions on Image Processing*, vol. 17, no. 5, pp. 664–678, May 2008. *Cited in Sec.* 3.1

[119] 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. *Cited in Sec.* (document), 1, 1.1, 2.2, 7.2

[120] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint arXiv:1703.10593*, 2017. *Cited in Sec.* 6.2.1, 6.3.1, 7.2