An Evaluation of HDR Image Matching under Extreme Illumination Changes

Aakanksha Rana, Giuseppe Valenzise, Frédéric Dufaux
CNRS LTCI, Telecom Paristech, Université Paris Saclay

***Abstract*—High dynamic range (HDR) imaging has potential to facilitate computer vision tasks such as image matching where lighting transformations hinder the matching performance. However, little has been done to quantify the gains with different possible HDR representations for vision algorithms like feature extraction. In this paper, we evaluate the performance of the full feature extraction pipeline, including detection and description, on ten different image representations: low dynamic range (LDR), seven different tone mapped (TM) HDR and two HDR imaging (linear and log encoded) representations. We measure the impact of using these different representations for feature matching using mean average precision (mAP) scores on four illumination change datasets. We perform feature extraction using four popular schemes in the literature: SIFT, SURF, BRISK, FREAK. With respect to previous studies, our observations confirm the advantages of HDR over conventional LDR imagery, and the fact that HDR linear values are not appropriate for vision tasks. However, HDR representations that work best for keypoint detection are not necessarily optimal when the full feature extraction is taken into account.**

***Index Terms*—High dynamic range imaging, tone mapping, descriptors, feature extraction, illumination change.**

I. Introduction

Many high-level computer vision algorithms based on local visual features such as object localization, tracking and classification, are extremely sensitive to changes in appearance of a scene in drastic illumination transformations. Even in mid-level tasks such as image matching, adverse lighting conditions can significantly worsen the performance of the feature descriptors [1]. High dynamic range (HDR) imaging [2] brings potential to surpass these limitations, thanks to its wider dynamic range which enables to capture details in both dark and bright regions.

Essentially, a feature extraction scheme consists of 2 parts: keypoint detection and descriptor computation. Feature extraction algorithms look for descriptors with the ability to describe the distinctiveness of the detected local regions in an image undergone different transformations (including lighting changes). Traditionally, these feature extraction algorithms [1] have been extensively explored with respect to low dynamic range (LDR) imagery, generally represented with a display-referred 8-bit integer representation. In contrast, HDR imagery consists of real-valued pixels proportional to the physical luminance of the scene, expressed in cd/m^2. Therefore, understanding which are the best modalities to apply LDR-based feature extraction techniques to HDR is an interesting and timely research question. For instance, HDR pixel values could be used directly, or could be firstly converted into a compact, 8-bit, displayable representation – an operation known as tone mapping (TM).

Some recent studies [3–7] have reported gains in terms of keypoint repeatability by using HDR-based imagery for keypoint detection, which is the primary block of feature extraction pipeline. These studies [3–7] have pointed that keypoint detection becomes unstable in low and high contrast areas of an LDR image. Similar advantages of HDR representations have also been shown for higher-level tasks such as tracking [8].

In summary, previous work either focused on evaluating detector performance only, with different HDR representations [3,5]; or, when the full feature extraction pipeline was evaluated, only a single HDR representation was considered [9]. However, a quantitative comparison of different possible HDR representations, including descriptors, has not been carried out so far. Therefore, it is difficult to draw precise conclusions on which is the best HDR representation (linear values, TMs, etc.) for an image matching pipeline, and how much is the gain with respect to LDR.

In this paper, we address these questions by conducting a performance evaluation of feature extraction algorithms for ten different HDR representations, including 7 TMs and 2 native HDR representations (linear and log-encoded) and standard LDR, using image datasets with drastic changes in illumination. The main novelty and contribution of this work is thus to consider the full feature extraction pipeline and measure the impact of different HDR representations on both detectors and descriptors. We compute the mean average precision performance for such a wide spectrum of image representations using 4 widely used feature extraction schemes: SIFT, SURF, FREAK and BRISK. We carry out the experimentations on publicly available datasets [3,4] with diverse lighting variations.

The paper is organized as follows: in Section II we provide the details of the evaluation setup. We present the experimental results and discussion in Section III and the conclusions in Section IV.

II. Evaluation Framework

This section is structured as follows: in Section II-A we highlight the HDR representations used for our evaluation; in Section II-B we briefly discuss the considered feature extraction schemes, followed by dataset selection in Section II-C; further, in Section II-D we detail the feature selection and matching strategy used for descriptors matching.

A. Image Representations

In this evaluation, we consider a total of 10 different image representations (listed in Table I) including the standard 8-bit LDR, 2 floating point HDR representations (HDRlog and HDRlin) and 7 different 8-bit tone mapping (TM) HDR representations. These TM techniques consist of 2 global and 5 local TMs. In general, *global* TM approaches are based on applying a compression function to all the pixels of the image, while *local* techniques computes tone-mapped pixels taking into account the values of neighboring pixels.

B. Feature extraction

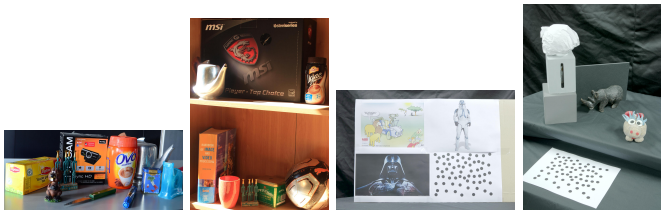
We assess the following 4 popular feature extraction schemes which employ both gradient-based histograms and computationally fast binary descriptors.

- **SIFT** [10]. This classic scheme is constituted of a blob keypoint detector (based on difference of Gaussians) and a gradient-based

The work presented in this document was upported by BPIFrance and Région Ile de France, in the framework of the FUI 18 Plein Phare project.

Abbreviations	Description
LDR	Best exposure LDR image of the scene
RNG(G)	A global scaled mapping operator [14]
DR(G)	An Adaptive logarithmic mapping [15]
RN(L)	A local dodging-and-burning operator [14]
MA(L)	Perceptual method for contrast processing [16]
FA(L)	Gradient domain HDR compression [17]
CH(L)	Spatially non-uniform scaling algorithm [18]
DU(L)	A fast bilateral filtering technique [19]
HDRLog	Logarithmic encoding in accordance to Weber-Fechner's law
HDRLin	Linear photometric luminance values stored in the HDR file

TABLE I: Different image representations for feature extraction.



Project Room Light Room 2D Lighting 3D Lighting

Fig. 1: Example images from the datasets employed in this work.

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** [11]. 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** [12]. 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.
- **FREAK** [13]. 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.

C. Datasets

We considered the following publicly available datasets:

- HDR illumination change datasets by Rana et al. [3] are composed of 2 parts: Project-Room with 8 lighting conditions and Light-Room with 7 lighting conditions as shown in Figure 1. These images have challenging lighting transformation scenarios with complex shaped objects, repeated patterns in texture, stark shadows and variety of illumination sources.
- 2D and 3D Lighting Dataset by Pribyl et al. [4] shown in Figure 1. It consists of 7 controlled lighting variations in each set. The 2D dataset is composed of a light-dark sectioned poster and the 3D dataset consists of few plain objects with fine geometry.

D. Keypoint Selection and Matching Metrics

Local feature extraction rely primarily on the detected keypoints and different detectors result in a different number of keypoints. Therefore, following the detection protocol by Rana et al. [3,4], we select 400 keypoints with the strongest detector response. In order to limit the detection in pertinent areas and ensure a fair comparison for feature extraction at later stages, we exclude keypoints from background and regions without meaningful objects as in [3,4]. In addition, we used the conventional detector parameters in [1,11,12] for LDR and TM image representations.

Next, for the descriptor part we compute standard precision-recall (PR) curves [1] for measuring the accuracy of matching. The PR curves are based on the number of true and false matches obtained for an image pair. A descriptor is said to have a match if it satisfies the nearest neighbor distance ratio (NNDR) criterion. According to NNDR, for a descriptor to find a good match, the ratio between its distance from first best match (dBM1) and its distance from second best match (dBM2) should be less than threshold th, i.e., $dBM1/dBM2 < th$. We use Euclidean distance for SIFT and SURF, Hamming distance for binary descriptors (FREAK and BRISK). Two descriptors yield a true positive match if they correspond to two keypoints which are indeed repeated [1] in the reference and query images. Similarly, a match is labeled as a false positive if the corresponding keypoints are not repeated. For PR curve computation, 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. By varying the NNDR th, we generate a PR curve and measure the area under the PR curve (AUC), also called as average precision (AP). The mean of APs for all image pairs is the mean average precision (mAP), which we have used to compare different representations for each extraction scheme.

III. 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). For each image pair, we compute the PR curve by varying th from 0.0 to 1.0 and record the AP value. After this, for each format and either feature extraction scheme, a mAP score is obtained by averaging the APs calculated on such 182 image pairs (see Table II). Higher mAP scores imply better descriptor matching.

Furthermore, to understand how detector and descriptor contribute to the overall performance, we expand our analysis to individual datasets and compute mAP and repeatability rate (RR) measures. Repeatability Rate is defined as the fraction of repeated points to the minimum number of detected points in the test or reference image. In Figure 2, 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 II show that HDRLin representation is consistently the worst performing using all extraction schemes. This is coherent with the previous findings in [3,5], and is mainly due to the low keypoint repeatability, which increases the probability of false positives. This leads to the first conclusion

[1] 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. $\epsilon = 5.0$ in our case

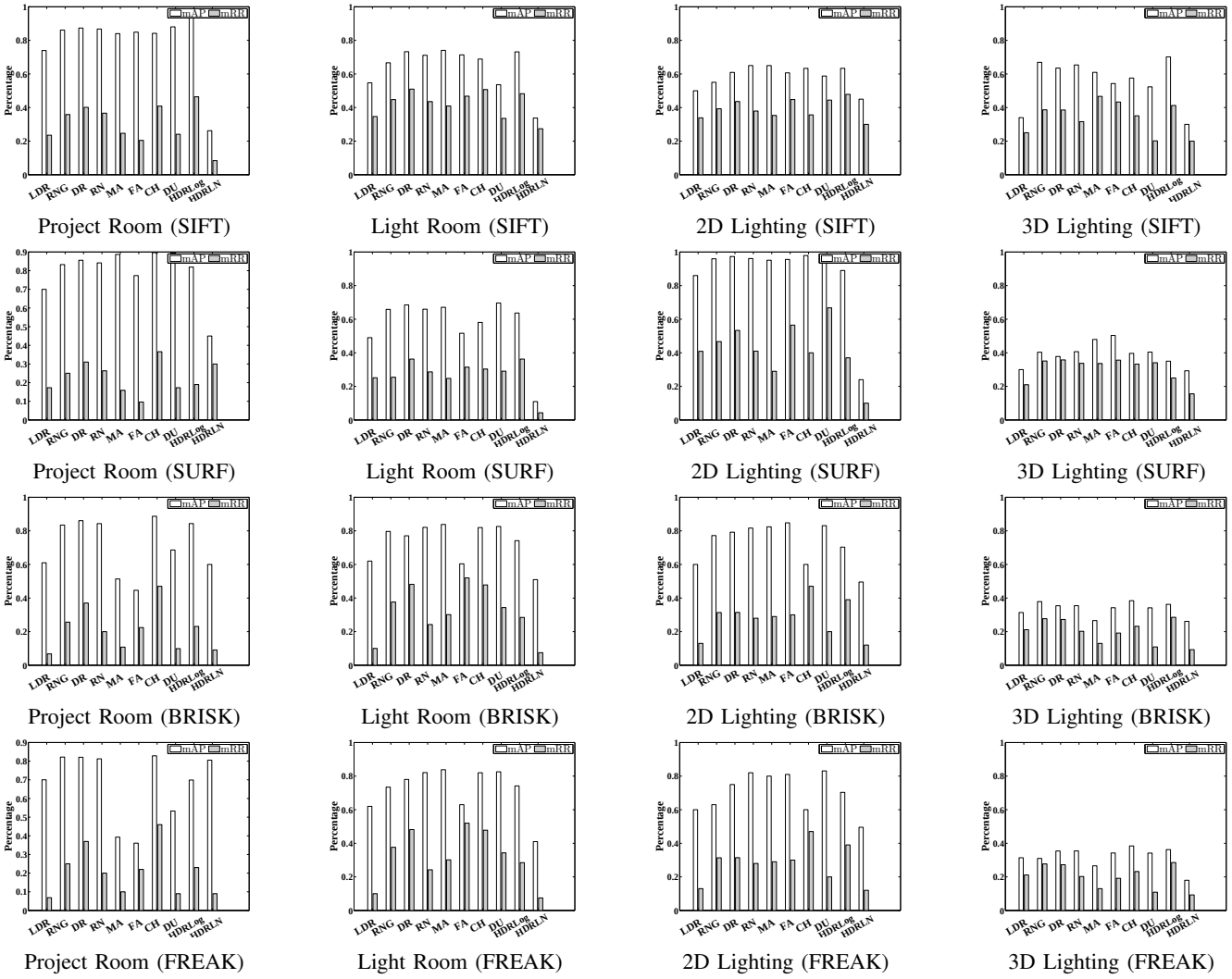
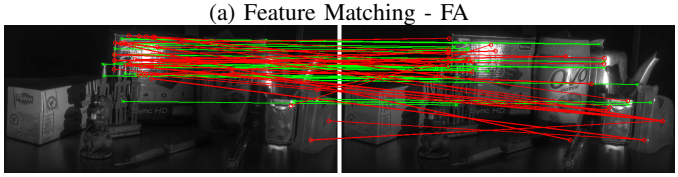


Fig. 2: Mean average precision (mAP) and mean repeatability rate (mRR) over the four considered datasets and feature schemes. mAP and mRR are computed on 56 image pairs, for the Project Room dataset, and over 42 image pairs for Light Room, 2D and 3D Lighting datasets.

(a) Feature Matching - FA



(b) Feature Matching - RN

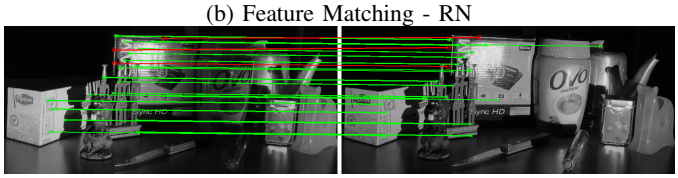


Fig. 3: An example of image matching for two TMs. The true positive and false positive matches are shown with green and red lines respectively. The TM in (a) achieves a higher repeatability (24 %) than that in (b); however, most of the matches in (a) are false positives, thus the AP for (b) is higher than in (a) (95 % vs. 87 %, respectively).

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 II: 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 double underlined.

of this work, i.e., HDRLin is not appropriate to be used for feature extraction algorithms, for both detector and descriptor.

HDRLog/TM versus LDR. On average, all HDR formats show significant gains of (at least) 8% mAP over single LDR exposure (see

Avg/Formats in Table II). 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 2, is the low repeatability rate which reduces the number of true positives.

HDRLog versus TMO's. mAP scores obtained from HDRLog and different TM formats 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 [20], as reported for keypoint repeatability in [3]. However, those representations require photometrically calibrated HDR pictures, which might not be available in practice.

Comparison with previous studies. Previous studies [3,5] have reported that local TM approaches such as Fattal or Chiu consistently provide more stable keypoints (in terms of repeatability) under illumination changes, compared to TMs 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 2 we observe that some TMs 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 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 2.

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.

IV. Conclusion

In this paper, we have presented a comprehensive evaluation of LDR and different HDR representations for image matching under lighting transformations. The analysis of mean average precision scores on different scenes confirms the potential of HDR tone mapping techniques over single LDR exposures. Furthermore, our study confirms that the linear high dynamic range values are inappropriate to be used for visual recognition tasks. More interestingly, we have also observed that local TMs with very high repeatability rate for feature detection are not necessarily the best option when the full feature extraction pipeline is considered. This suggests 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.

References

[1] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

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

[3] 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.

[4] P. Bronislav, A. Chalmers, and P. Zemčík, "Feature point detection under extreme lighting conditions," in *Spring Conference on Computer Graphics*, 2012, pp. 156–163.

[5] P. Bronislav, A. Chalmers, P. Zemčík, Lucy Hooberman, and Martin Čadík, "Evaluation of feature point detection in high dynamic range imagery," *Journal of Visual Communication and Image Representation*, vol. 38, pp. 141 – 160, 2016.

[6] G. Kontogianni, E. K. Stathopoulou, A. Georgopoulos, and A. Doulamis, "HDR imaging for feature detection on detailed architectural scenes," *Int. Journ. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 325–330, 2015.

[7] 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, Sept. 2016.

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

[9] L. Chermak and N. Aouf, "Enhanced feature detection and matching under extreme illumination conditions with a hdr imaging sensor," in *IEEE 11th Int. Conf. on Cybernetic Intelligent Systems*, Aug 2012, pp. 64–69.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[11] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *In ECCV*, pp. 404–417. 2006.

[12] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *Proceedings of the 2011 International Conference on Computer Vision*, Washington, DC, USA, 2011, ICCV '11, pp. 2548–2555.

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

[14] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graph.*, pp. 267–276, July 2002.

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

[16] R. Mantiuk, K. Myszkowski, and H. P. Seidel, "A perceptual framework for contrast processing of high dynamic range images," *ACM Trans. Appl. Percept.*, vol. 3, no. 3, pp. 286–308, July 2006.

[17] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 249–256, July 2002.

[18] 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*, Toronto, Ontario, Canada, 1993, GI '93, pp. 245–253.

[19] 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*, 2002, SIGGRAPH '02, pp. 257–266.

[20] T. Aydın, R. Mantiuk, and H.P. Seidel, "Extending quality metrics to full dynamic range images," in *Human Vision and Electronic Imaging XIII*, January 2008, Proceedings of SPIE, pp. 6806–10.