Evaluation of Bright and Dark Details in HDR Scenes: A Multitask CNN Approach

Gabriel Pacianotto, Daniela Carfora, Franck Xu, Sira Ferradans, Benoit Pochon. DXOMARK

Abstract

High dynamic range (HDR) scenes are known to be challenging for most cameras. The most common artifacts associated with bad HDR scene rendition are clipped bright areas and noisy dark regions, rendering the images unnatural and unpleasing. This paper introduces a novel methodology for automating the perceptual evaluation of detail rendition in these extreme regions of the histogram for images that portray natural scenes. The key contributions include 1) the construction of a robust database in Just Objectionable Distance (JOD) scores, incorporating annotator outlier detection 2) the introduction of a Multitask Convolutional Neural Network (CNN) model that effectively addresses the diverse context and region-of-interest challenges inherent in natural scenes. Our experimental evaluation demonstrates that our approach strongly aligns with human evaluations. The adaptability of our model positions it as a valuable tool for ensuring consistent camera performance evaluation, contributing to the continuous evolution of smartphone technologies.

Introduction

The Human Visual System (HVS) can adapt to very extreme light conditions that can expand a dynamic range of 120 dB with adaptation, and up to 40 dB without adaptation [17]. However, most cameras find it challenging to reproduce fine details under High Dynamic Range (HDR) scenes (with dynamic range up to 80 dB). The difficulty lies in their ability to properly expose both dark and bright areas simultaneously and render (tone map) the information pleasantly into a limited dynamic range. This problem often results in clipped bright lights (saturated pixels that are fully white) and/or shadow detail loss (very dark or noisy pixels as the sensor's response falls below its noise threshold), leading to the loss or alteration of visual information.

Given the current complexity of the smartphone ISP pipeline and the importance of its tuning, automatic Image Quality Assessment (IQA) has gained significant importance. IQA serves as a critical tool in evaluating the performance of image processing algorithms, ensuring that they effectively capture and represent scenes with varying levels of brightness and contrast, thereby enhancing overall image quality.

Our research focuses on the perceptual evaluation of bright preservation (BP) and dark recovery (DR) attributes, highlighting the crucial role of dynamic range in determining image quality.

Contributions

This study encompasses several key contributions:

1) The construction of a robust BP and DR dataset, based on pairwise perceptual evaluations. Additionally, we introduced a post-experiment control of the annotations to identify and fix any outliers that are not aligned with the established guidelines.
2) The development and training of no-reference image quality models for dynamic range quality assessment. Notably, our approach distinguishes itself by acknowledging the different, yet complementary, perceptual evaluations of bright preservation and dark recovery. To achieve this, our approach considers the unique regions of interest across diverse scenarios, something that has not been addressed before in the IQA field.

Related work

A good image quality rendition in natural scenes, particularly those with a large illumination range, presents a significant challenge for devices during the capture process. Few previous studies have explored both Image Quality Assessment (IQA), and High Dynamic Range (HDR) image rendering to address these challenges.

IQA can be broadly categorized into Full-Reference (FR) and No-Reference (NR). FR-IQA compares a reference and a distorted/test image to predict the perceptual quality of the test image. In contrast, NR-IQA methods output an absolute quality measure without taking into account any reference. They are particularly relevant in real-world scenarios where obtaining a pristine reference image may be impractical or impossible.

Full-Reference IQA. Most works on FR-IQA for bright and dark detail rendition have been proposed in the context of quantifying the amount of distortion introduced by tone mapping operators, that is, algorithms that allow to visualize HDR images in low-dynamic screens. Aydin et al. [3], was the first to introduce an image quality metric inspired by the HVS capable of comparing images with different dynamic ranges. Their metric identifies and detects 3 different common distortions introduced by a tone compression operator (“loss of visible contrast”, “amplification of visible contrast”, “reversal of visible contrast”). Also, for tone mapping quality assessment, Yeganeh and Wang [18] proposed the *Tone Mapped image quality index*, which is an extension of the Structural Similarity Index to the case where the reference and the evaluated image do not have the same dynamic range. Kundu et al [19] extended this method by considering saliency models. Song Y, et al, [6] introduced context-region assessment by proposing a quality metric that also considers the color distortion. However, objective measures are known to correlate poorly with human perception [20].

No-Reference IQA. Since the original image is usually unavailable for use as a reference in many applications, it is necessary to move into the NR-IQA. Most NR-IQA methods for evaluating images from HDR scenes target HDR format images [7,10,11,12], that is, images that need to be displayed on HDR screens. But this scope, in terms of image type coverage, is relatively small given the current cost and presence of HDR

displays in the consumer market. Here, we are interested in analyzing low dynamic range images, independently of the process used to generate them. There is a notable gap in the availability of a robust perceptual database founded on pairwise comparisons, tailored for both Bright Preservation and Dark Recovery aspects since existing research predominantly concentrates on assessing overall image quality in both SDR and HDR images, neglecting a specific emphasis on preserving intricate details within the bright and dark regions.

Proposed method

We propose a different approach to NR-IQA for low dynamic range images. Our goal is to estimate the quality of a photograph from a predefined scene known to be HDR. This is a task that humans can do very easily (less than 5 seconds). To this end, we gathered a quality-diverse and precisely annotated dataset of images, to serve as ground truth for training a multitask machine learning algorithm. The numerical results show the relevance of this approach.

Dataset

The database is composed of 25 scenes for Bright Preservation and 25 scenes for Dark Recovery. A scene here denotes a set of images with similar content captured in the same location with similar viewing angles. We consider portrait and non-portrait scenes, as can be seen in Figure 1. For each scene, we have at least 100 images resulting in a database of about 2500 images for Bright Preservation and 2500 images for Dark Recovery. The images are sourced from over 148 devices, guaranteeing a thorough representation of the market's quality range.

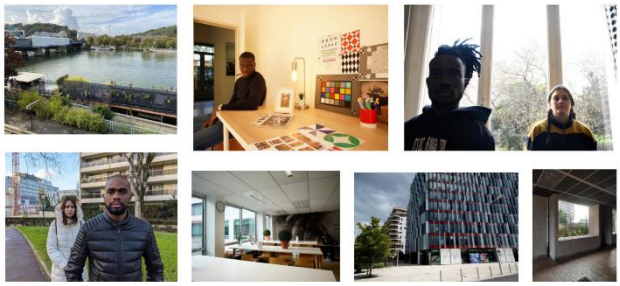

Figure 1 *- Scene examples for BP and DR datasets. Portrait and non-portrait.*

Annotation strategy

For all 50 scenes in our database, images within a scene have been perceptually annotated using a pairwise comparison (PWC) methodology. Each scene was annotated independently.

The annotation task was done under controlled conditions, with no direct illumination on the screen. The displays were calibrated (D65 white point with peak luminance at 120cd/m2). The viewing condition was also fixed: the images were displayed side by side at a distance to the eye of 65cm, on a 32'' 16:9, UHD 4K screen. Annotators were able to zoom in to see both images simultaneously at 1:1.

The opinion of more than 20 experts was gathered using an internal PWC tool. Observers were asked to select the best out of two images, following predefined guidelines. This process involved conducting up to 1.0 standard trials per scene. Each standard trial involved evaluating n(n−1)/2 pairs for an n-image set.

To optimize the cost of the pairwise comparison task, Active Sampling (ASAP), as referenced in [4], was employed. This strategy prioritizes the selection of image pairs that offer the most valuable information. By acknowledging that not all comparisons are equally useful, the Active Sampling method enhances the efficiency of the evaluation process.

Psychometric scaling. Designing a PWC experiment requires modeling the statistical distribution of the image quality. Based on the Thurstone Case V observer model, outlined by Perez-Ortiz et al.[1], the quality of an image is described as a Gaussian distribution $\boldsymbol{N(\mu, \sigma)}$. The average μ represents the actual quality and $\boldsymbol{\sigma^2}$ is its "perceptual" variance across observers.

The results are typically expressed in *Just-Objectionable-Difference (JOD)* units. Two images are 1 JOD apart if 75% of observers choose one as better than the other, and a random guess (i.e., probability of 50%) results in a JOD distance of 0 between the images.

Annotator outlier analysis. In our outlier analysis study, we employed the CrowdBT model [2]. This model assumes equal treatment of each annotator and introduces the *CrowdBT coefficient* defined as:

$$\boldsymbol{\eta_k = Pr\left(X_i \succ_k X_j \,\middle|\, X_i \succ X_j \right)} \qquad (1)$$

where η_k represents the probability that annotator k chooses the image i over the image j, assuming a perfect annotator would make the same choice.

We calculated the CrowdBT coefficients η_k using the Maximum Likelihood Estimator. For an ideal annotator ($\eta_k \approx 1$), spammer ($\eta_k \approx 0.5$), and malicious or poorly informed annotator ($\eta_k \approx 0$), distinct values were obtained.

Figure 2 illustrates the pattern of annotators (designated by capital letters) across different scenes. It highlights significant variations in behavior both among annotators and across scenes. Notably, annotators like 'F' who consistently rate 0 (malicious) with a mean score below 0.5 (indicative of spamming) can disrupt the integrity of the JOD scale. Therefore, it is crucial to identify and address such outliers to ensure the reliability of the final scale.

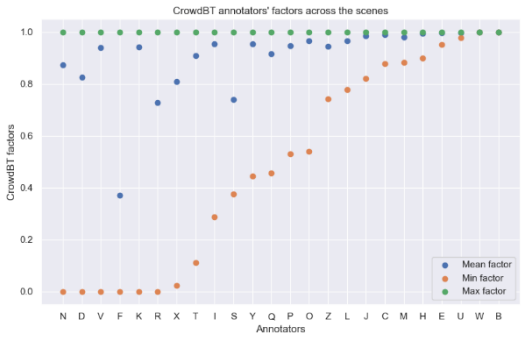

Figure 2 *- CrowdBT annotators' factors across the scenes. Mean, min and max correspond to the annotator's computed crowdBT factors across all scenes.*

When $\eta_k \approx 0$ the annotator inverted most of the comparisons. In such a case, we decided to invert the annotator comparison matrix instead of discarding it. This approach helps maintain the number of total comparisons, very important to avoid poor convergence of the quality scale.

This methodology proved valuable in identifying annotator patterns and refining our annotations. The final ground truth is a set of images, structured by scenes, where every image has a robust JOD score assigned. Note that, by construction, each scene has an independent quality scale.

Machine learning model for quality prediction

We implement a multitask learning approach where the prediction of the quality score is supported by the classification of the scene type (25 scenes for Bright Preservation and 25 scenes for Dark Recovery) as an auxiliary task. The combination of the two tasks allows us to train on all images, independently of their scale, making the model capable of generalizing across different scenarios and shooting conditions.

As can be seen in Figure 3, the only input of the model is the complete resized image. We use the backbone of a pretrained ResNet-50 [14] to extract relevant features from the image. These extracted features from the input image are then channeled into two distinct customized fully connected heads: one for quality prediction and the other for scene class prediction, as described in [15]. This results in the prediction of a float value for the quality and an integer value mapping to the scene type. Implementation details can be found in Appendix A

JOD score to AI global score mapping

During the annotation process, each scene was annotated separately to ensure the high quality of the final annotations, given that cross-content image comparison is prone to inaccuracies. Therefore, by construction, the JOD quality scores are only meaningful within each content group (i.e., scene).

Each scene has a different JOD range that is linked to its intrinsic difficulty:

- For an easy scene, the difference between a good and a bad device is less visible for the annotators, and therefore the JOD range is smaller.
- For a challenging scene, the difference between a good and a bad device is important, with many possible intermediate cases. Therefore, we expect a larger JOD range.

This means that comparing the JOD scores across different scenes is not a straightforward task. If we want to, for example, compare two devices shot across multiple scenes, we need a way to aggregate and compare scores from different scenes. To address this issue, it is crucial to map the JOD scale into a comparable score. This is achieved by adding a final block at the end of the model prediction pipeline. This block applies a function taking as inputs the quality score in JOD and the scene class prediction, producing an output score ranging from 0% to 100%. A diagram of the complete framework can be seen in Figure 3.

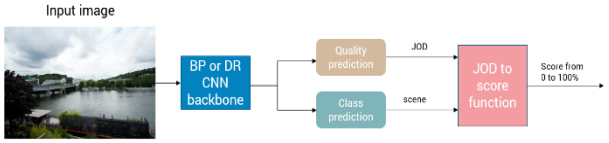

Figure 3 *- Diagram of final model: A multitask CNN that predicts the scene class and the JOD image quality. A final block maps the JOD and the scene prediction into a cross-scene comparable score, see text for more details.*

The choice of the mapping function needs to meet some constraints:

- Monotonically non-decreasing function: The score should consistently increase with higher JOD values.
- Bounded: The score should align with the human perception evaluation.

Given those constraints, a Sigmoid function is chosen. The parameters of the Sigmoid were empirically fine-tuned for each scene according to its difficulty.

It is important to note that this final block has fixed weights chosen empirically, which are not changed during the training of the model itself.

Experiments

Dataset Split

Before starting the training process, the annotations were split into 3 different datasets: Train (60%), Validation (20%) and Test (20%). Splitting the dataset is a very important task as unbalanced sets might lead to biased predictions and poor results. The description of the optimization problem and the considered constraints can be found in Appendix B.

Model performance

Performance was evaluated with 3 metrics:

- Mean Average Error (MAE). The smaller the better.
- Spearman's Rank correlation coefficient (SROCC): The bigger the better
- Pearson's linear correlation coefficient (PLCC): The bigger the better.

Metrics were computed on each scene separately and the mean over all the scenes is reported in Table 1. The model shows a strong correlation with human perceptual evaluations and an MAE inside the interval in which a person, on average, cannot distinguish a change in quality, that is, 1 JOD.

MODEL	MAE (JOD)	SROCC	PLCC
BRIGHT PRESERVATION	0.55	0.88	0.90
DARK RECOVERY	0.59	0.86	0.86

Table 1 *- Performance of Bright Preservation (BP) and Dark Recovery (DR) models.*

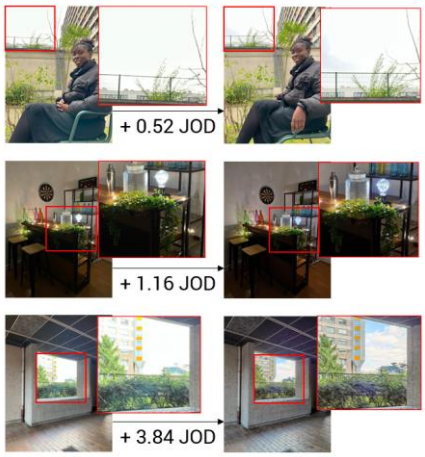

Figure 4 *- Example of repeatability tests performed with the Bright Preservation model. Images in the same row were taken with the same camera in different shooting sessions. See text for more details.*

Repeatability tests

A common problem seen on current cameras is their stability. Photos captured one after the other by the same camera may present different renderings, due to the trigger of different camera treatments. An important task the proposed models should be able to handle is to identify those instabilities, whether they are small or big.

To check the capacity of the proposed models to identify repeatability problems, images coming from the same shooting session were passed through the models and their results were perceptually checked. It is important to notice that those images were not included in the original dataset, and represent images never seen before by the models. In Figure 4 and Figure 5, we can see 3 sets of images ordered by rows. On the left, we see the worst image of the same shooting sequence, and on the right the best. First, we see that the differences are clearly visible in the bright areas (Figure 4) and the dark regions (Figure 5). As expected, we can see that both the Bright Preservation and the Dark Recovery model are able to identify and quantify repeatability problems. This is true for both small differences (0.52 and 1.16 on Figure 4, or 0.55 and 0.99 JOD on Figure 5)

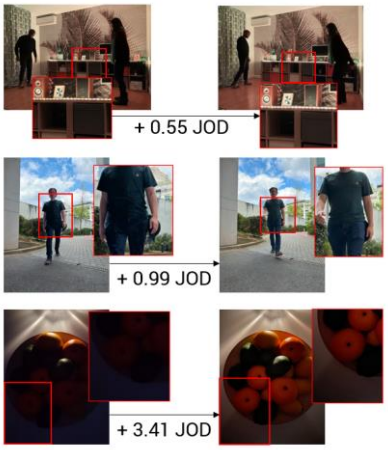

Figure 5 *- Example of repeatability tests performed with the Dark Recovery model. Images in the same row were taken with the same camera in different shooting sessions See text for more details.*

and big differences (JOD bigger than 3 on the last row of both figures).

Comparison of different device shooting sessions

Weather conditions, the hour of the day, or even the season can have a great impact on a (smartphone) camera performance and therefore, the perceptual analysis of its images might give different results depending on the different shooting sessions. Moreover, it is well known that camera processing systems may present instabilities that make devices behave differently.

Therefore, to understand the quality changes related to the processing instability from the quality changes due to the variation of the natural scene (less light due to the shooting hour, for example), we compare our target camera with other cameras of similar quality, expecting the quality of their images to correlate to our target camera across shooting sessions if they are all stable. As an example of this, two devices were shot in 5 different shooting sessions. Our BP CNN model was used on every image, and the mapping function was used to transform the JOD inference into a score. Scores were aggregated by doing an average over every relevant scene (same lighting condition and same scene use case, in this case, Portrait/non-portrait). Figure 6 shows an example of the Outdoor Portrait Bright preservation score.

Two things can be concluded from the graph:

- Scores are dependent of shooting session, which confirms the need of reference devices.
- Scores of different devices are correlated over different shooting sessions.

Figure 7 shows some of the scenes used to compute those scores for device A on shooting sessions 2 and 5. The shooting

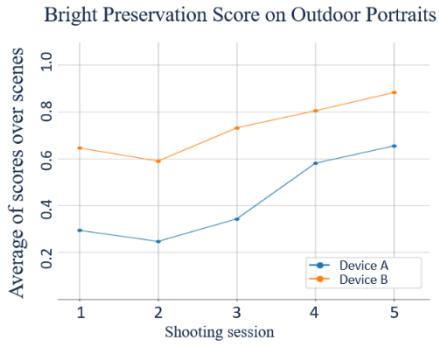

Figure 6 *- Brights preservation Outdoor Portraits score of two devices over different shooting sessions.*

session 2 was effectively harder for device A, as the sky is strongly clipped in that session, which matches the conclusion from the graph.

A similar study done for the DR model can be found in Appendix C.

Figure 7 *- Example of scenes of Device A over shooting sessions 2 and 5*

Conclusion

This paper introduces two models that can evaluate the Brights Preservation and Darks Recovery of 25 scenes each. To do so, over 5000 images were annotated by 20 annotators on a pairwise comparison annotation campaign. During the campaign, an active sampling algorithm was used to optimize the comparisons. This dataset was used for training two multitask CNNs that obtained performing results when evaluated on a device-split testing set (cameras that were on the training set were not present in the testing set). Moreover, the provided JOD scores were converted to fix-range scores that make them comparable across scenes.

The proposed models are shown to correctly predict the bright preservation (BP) and dark recovery (DR) attributes. They also serve to identify the difficulty of the shooting session on which the pictures were captured, which can potentially save a lot of resources when comparing different shooting sessions, and reduce the bias introduced by instabilities related to the shooting conditions.

References

[1] M Perez-Ortiz and R K Mantiuk. “A practical guide and software for analyzing pairwise comparison experiments”. 2017.

[2] Xi Chen and Paul N Bennett. “Pairwise Ranking Aggregation in a Crowdsourced Setting”. In: Proceedings of the Sixth ACM International Conf. on Web Search and Data Mining. WSDM ’13. Rome, Italy: Assoc. for Computing Machinery, 2013, pp. 193–202.

[3] Aydin, T., Mantiuk, R., Myszkowski, K., Seidel, H. 2008. Dynamic Range Independent Image Quality Assessment. ACM Trans. Graph. 27, 3, Article 69 (August 2008), 10 pages. DOI = 10.1145/1360612.1360668 http://doi.acm.org/10.1145/1360612.1360668

[4] A. Mikhailiuk, C. Wilmot, M. Perez-Ortiz, D. Yue and R. K. Mantiuk, 2020. “Active Sampling for Pairwise Comparisons via Approximate Message Passing and Information Gain Maximization”, International Conf on Pattern Recognition (ICPR)

[5] Narwaria, M., Perreira Da Silvam M., Le Callet, P., “Study of High Dynamic Range Video Quality Assessment”. 2016.

[6] Yang Song, Gangyi Jiang, Mei Yu, Zongju Peng, Fen Chen, “Quality assessment method based on exposure condition analysis for tone-mapped high-dynamic-range images”. Signal Processing., 2017, doi: 10.1016/j.sigpro.2017.12.020

[7] Fan K., Liang, J., Li, F. and QIU, P., “CNN Based No-Reference HDR Image Quality Assessment”. Chinese Journal of Electronics Vol. 29, No 2, 2021.

[10] Gangyi Jiang et al. “No-reference quality metric for high dynamic range imaging system based on curvature analysis in tensor domain”. Optica Applicata, Vol. XLIX, No. 4, 2019. DOI: 10.37190/oa190401

[11] Kottayil et al. “Blind Quality Estimation by Disentangling Perceptual and Noisy Features in High Dynamic Range Images” IEEE Transactions on Image Processing, Vol 21, No. 3, 2018.

[12] D. Kundu and B. L. Evans et al. “No-Reference Image Quality Assessment for High Dynamic Range Images” IEE, 2016.

[13] Prasetya G., et al. “A review on high dynamic range (HDR) image quality assessment”. International Journal on Smart Sensing and Intelligent Systems. Issue 1, Vol. 14, 2021

[14] He, Kaiming & Zhang, Xiangyu & Ren, Shaoqing & Sun, Jian. (2016). Deep Residual Learning for Image Recognition. 770-778. 10.1109/CVPR.2016.90.

[15] Chen-Hsiu Huang and Ja-Ling Wu. Multi-task deep cnn model for no-reference image quality assessment on smartphone camera photos, 2020.

[16] Kantorovich, L. V. Mathematical Methods of Organizing and planning Production. Management Science. 6 (4): 366{422. doi:10.1287/mnsc.6.4.366. JSTOR 2627082 (1939).

[17] Darmont, Arnaud. "High dynamic range imaging: sensors and architectures." *SPIE* (2013).

[18] 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

[19] Kundu, Debarati, et al. "No-reference quality assessment of tone-mapped HDR pictures." *IEEE Transactions on Image Processing* 26.6 (2017): 2957-2971.

[20] I. R. Khan, T. A. Alotaibi, A. Siddiq and F. Bourennani, "Evaluating quantitative metrics of tone-mapped images", IEEE Trans. Image Process., vol. 31, pp. 1751-1760, 2022.

Acknowledgments

This project was provided with computer and storage resources by GENCI at IDRIS thanks to the grant 2023-AD011014305 on the supercomputer Jean Zay's the V100.

Author Biography

Gabriel Pacianotto Gouveia *holds a double degree from Escola Politécnica da Universidade de São Paulo (Poli-USP, Brazil) and Ecole Centrale Paris (France), with a specialization in Electrical Engineering. He joined DXOMARK in 2019 as an Image Quality Engineer, and since 2022 he works as a Machine Learning Engineer, helping to develop new image assessment techniques with the use of machine learning.*

Daniela Carfora Ventura *is a machine learning engineer, with a focus on the image domain, at DXOMARK. She holds a double degree from Simón Bolívar University (Caracas, Venezuela) and Télécom SudParis, with a Master's degree in Data Analysis and Pattern Classification. She actively contributes to the end-to-end machine learning pipeline for image quality assessment.*

Franck Xu *received his Engineering Degree with specialization in Signal Processing and Machine Learning at Institut Mines-Télécom (IMT) Atlantique (France) in 2023.*

Sira Ferradans *is currently the AI director at DXOMARK. She has earned her PhD in Computer Vision from the Universitat Pompeu Fabra (Barcelona, Spain), and worked as a researcher at Duke University (North Carolina, US) and Ecole Normale Superieur (ENS Paris, France). Since 2016, she works in the industry bridging the gap between research and product in the machine learning domain.*

Benoit Pochon *received his Master’s degree in engineering from Centrale Supelec (2001) and his Master’s degree in Electrical Engineering from GeorgiaTech University (2001). After several years working in the signal processing domain, he joined DXOMARK Image labs in 2017, as image science director.*

Appendix A - Implementation details

	REGRESSION LOSS WEIGHT	CLASSIFICATION LOSS WEIGHT
BP	0.9	0.1
DR	0.8	0.2

Table 2 - *Different loss weights for each model*

The training was done using a 32Gb Nvidia V100, using a batch size of 20. Each input image was resized to 600x600, using LANCSOS interpolation. Adam optimizer was employed with an initial learning rate of 1e-4 and a weight decay of 5e-4. To adjust the learning rate during training, a scheduler was implemented with a gamma of 0.9 and a step size of 10. Early stopping was activated by monitoring the validation SROCC.

The loss is a weighted average of the Huber loss for the regression and Cross-entropy for the classification, ensuring proficiency in both tasks. The weights were chosen empirically, and the weights that gave better results are shown in Table 2.

Appendix B – Dataset split

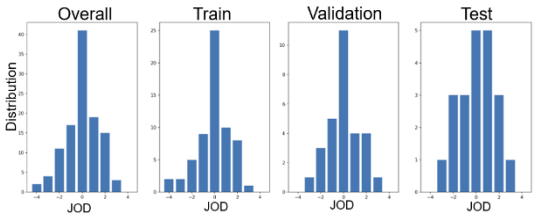

Figure 8 *- Example of JOD histogram for the 4 datasets for a given scene: overall, train, validation, and test*

To perform the split, the following constraints were considered:

- *Split per device:* all the images taken by one device, or similar devices (same brand and same camera hardware) belong to one set no matter the scene. That is to say, we cannot find an image taken with the same device in the train and the test set.
- *Split per quality:* all sets should have a similar JOD distribution for each scene.

Splitting the dataset can be modeled as an optimization problem: we are trying to maximize the similarity of the distribution of the 3 datasets to the complete dataset, for every scene, given the device constraints described above. All the constraints should be taken into account at the same time, as we want a single dataset split in the end and not a split per constraint. To estimate the similarities between the dataset distributions we used the Earth's mover distance, also known as Wasserstein metric [16].

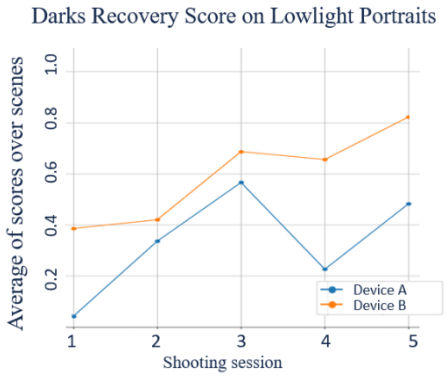

Figure 9 *- Darks Recovery Lowlight Portraits score of two devices over different shooting sessions.*

The complete dataset represented 25 scenes for BP and 25 for DR, coming from 148 different devices which were categorized into 83 device categories. Those categories were randomly sampled and put into one of the 3 sets, and then the chosen metric was computed. This experiment was repeated 100000 times, and the experiment with the smaller Wasserstein metric was chosen as our dataset split. An example of the distribution of a scene can be seen in Figure 8.

Appendix C - Comparison of different device shooting sessions for Darks Recovery

Figure 10 *- Example of scenes of Device A and B over shooting sessions 1 and 3*

As for the study performed on Brights Preservation, 2 devices were shot on 5 different shooting sessions. The scores for Darks Recovery can be found in Figure 9. This time, we can see that the score correlation was not as evident as in the previous case. Figure 10 shows one of the scenes used to compute the score, for devices A and B for both the shooting sessions 1 and 3. We can clearly see that device A severely underexposed the scene on shooting session 1, which explains the big dip in the score for that device and that shooting session. This points then to the fact that different score tendencies for devices over different shooting plans can be explained by the device's unstable pipelines.