https://doi.org/10.2352/EI.2026.38.10.HVEI-238

On a Perceptually Accurate Visual Noise Metric for HDR Imaging: Accounting for Luminance Adaptation and Gradient Effects

Hugo Masson, François-Xavier Thomas, Claudio Greco, Daniela Carfora Ventura, Mauro Patti, Benoît Pochon, Hoang-Phi Nguyen, Laurent Chanas, Frédéric Guichard
DXOMARK, Boulogne-Billancourt, France

Abstract
Currently available visual noise estimation algorithms are primarily developed and calibrated using SDR images, which limits their accuracy in representing the actual noise perceived by humans in HDR content. One key factor often overlooked is the luminance adaptation of the observer, especially when there is a significant contrast between the observed patch and its surrounding area. Moreover, the design of existing test charts, combined with increasingly sophisticated local tone mapping algorithms, introduces new challenges. A prominent issue is the presence of gradients in the final image, which significantly affect algorithmic measurements but have minimal impact on human perception: for instance, a patch may register a Just Noticeable Difference (JND) of 6 compared to a perfectly clean patch with zero noise, despite being visually clean and not visibly different. This paper proposes a new direction for visual noise algorithms and sets the foundations for future research. It presents findings on: 1) A new HDR ruler for visual noise assessment. 2) The impact of various factors (CSF, HPF, gradient correction) on algorithm performance. 3) Evaluation of different color spaces to calculate visual noise metrics.

Introduction
Visual noise metrics aim to predict the reaction of a human observer evaluating the presence of noise in a uniform region of an image (a "patch", in synthetic lab scenes) with a single scalar metric representing how much noise they see, or "noisiness". These metrics are particularly useful compared to others (*e.g.*, SNR) thanks to their approximate perceptual uniformity.

In the following, we want to address two limitations we found in all existing approaches:

- we encountered unexpected results when testing images captured by real devices where the perception did not correspond to the computed metric, notably because of low-frequency non-uniformities ;
- the existing algorithms were tuned on SDR luminance ranges and colorspaces which make evaluating HDR media more difficult.

In the sections below, "2021 dataset" is the dataset from [1], "2025 dataset" is the new dataset created in this paper ; `DXOMARK2021` [1], `CPIQ2023` [2] and `ISO2023` [3] are the existing Visual Noise metrics.

Initial observations and hypotheses
Observation 1: Non-uniformities affect existing Visual Noise metrics

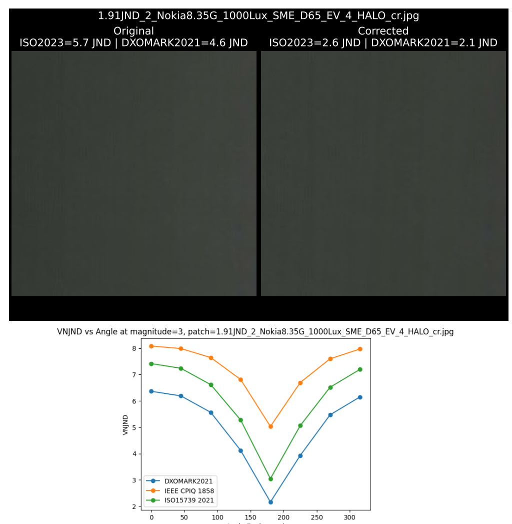

Figure 1. *Patch from the 2021 dataset before (left) and after (right) gradient correction described in algorithm 1. Even in previous experiments, this patch was a clear outlier for all Visual Noise metrics. The evolution of existing Visual Noise metrics (bottom) was computed on the same patch after synthetically modifying it with a multiplicative factor in the form of a linear gradient with the direction given by an angle.*

The two patches on the top of figure 1 appear very similar visually, but all Visual Noise metrics predict a significant difference of more than 2 JND. This is not an isolated case and matches several cases we encountered during out testing: to show this, we measured existing Visual Noise metrics on images from the JND ruler from the 2021 dataset, modified by overlaying a multiplicative linear luminance gradient defined by a slope (intensity) and an angle (orientation). We observed a consistent trend: for the vast majority of patches, there exists at least one gradient configuration that significantly minimizes the JND output (thereby correcting an existing non-uniformity), often by a large factor (such as for example in the graph in the same figure).

In the case of HDR media on HDR scenes, the presence of non-uniformities could originate from either veiling glare or from the local tone mapping process. Although this variation is often negligible to human observers, it produces a large variance in the output of the existing models, which is interpreted as noise. For this reason, Visual Noise ISO and CPIQ standards require patches to be free from non-uniformities as much as possible, but the prevalence of this issue makes them unusable on a large number of real camera images nowadays.

To better evaluate their impact, both on human perception and on algorithmic assessments, we conducted a quick pilot experiment. A series of noisy patches with non-uniformities are corrected with algorithm 1, and two randomized groups of images are formed: each uncorrected patch is randomly assigned to one group, with its gradient-corrected counterpart placed in the other group.

Participants are then randomly assigned to one group only (no participant performed both, thus two different observer populations were tested), and asked to rank a series of patches according to their perceived "noisiness".

Viewing distance was not fixed, since relative comparisons apply equally to all patches. The experiment was performed on an Apple XDR Pro display, calibrated to the ISO 22028-5 [4] profile, using the Google Chrome browser.

Even in this limited experiment, both observer groups provided consistent rankings between gradient-corrected and uncorrected patches (figure 2b). In contrast, the algorithms performed inconsistently and were strongly affected by the presence of gradients (figure 2a).

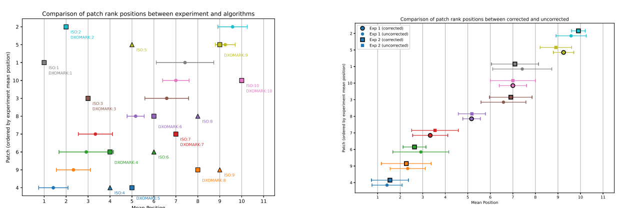

Figure 2. Two graphs presenting the results of the gradient experiment: (a) on the left, observer rankings of original patches without correction (small circles) are compared to the predictions of the ISO (squares) and DXOMARK 2021 (triangles) Visual Noise Metrics ; (b) on the right, the same original observer rankings (small circles) are compared to observer rankings of corrected patches (squares) following the procedure described in algorithm 1.

Some Visual Noise metrics also use low-pass filters to eliminate low-frequency variations (figure 3), but they are in light of these results clearly not sufficient. This also highlights a crucial question: how should we distinguish between what is considered "noise" and what is considered a "low-frequency variation", discounted by observers? This boundary may vary from one observer to another and change depending on viewing conditions: it is difficult to verify experimentally for now.

Observation 2: Observer adaptation is not taken into account by existing Visual Noise metrics

A first observation is that Visual Noise metrics are based on the CIELAB colorspace, which takes as input the measured stimulus XYZ, but also a normalization "white point" stimulus $X_n Y_n Z_n$; for example (with f being the CIELAB nonlinear mapping function) $L^* = 116 f(Y/Y_n) - 16$.

SDR displays with a limited range are typically assumed to reproduce printed media, and as such adopt a "white" color emitted by the display at its peak intensity as the CIELAB normalization white point, which in this restricted case can be a reasonable description of the observer's adaptation state. In the general case (and notably HDR media), this is not necessarily the case, and a potential mapping from observer adaptation to a single stimulus is undetermined. Even in SDR media, the case of a sufficiently dark image viewed in an equally dark room is not well-supported by existing metrics: many images will appear as noisy despite the L^* values, variances and predicted JND values being very small.

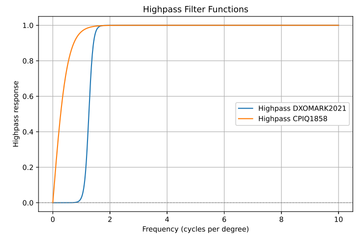

Figure 3. `DXOMARK2021` and `CPIQ2023` high-pass filter response on the same plot, showcasing the difference and the implicit choices both have made.

A second observation is that observer adaptation is an important factor in nearly all Color Appearance Models (among others: [5], [6], [7], [8]). It would seem odd that a Visual Noise metric does not take this into account, particularly for HDR media with large luminance contrasts. An attempt at using a weighting function based on average L^* to model adaptation was introduced in `DXOMARK2021` ; however, the CIELAB white point already handles this to some extent.

Finally, prior to this, all the available Visual Noise datasets were SDR media with a limited luminance range, annotated by observers in a fixed adaptation state (driven by the an approx. 20% grey background used in these experiments). This makes it more difficult to evaluate the models' accuracy in different adaptation states.

Our initial hypothesis following anecdotal observations is that the further the displayed luminances are from the adaptation point (to a first approximation: close to the same luminance as the background), the harder it is to discriminate noise.

Full Pairwise experiment

To better understand that, we conducted a more complete psychovisual experiment, whose major goal was compatibility with HDR media. To that effect, we use an HDR display with a peak of 1000 nits, as something that is now relatively commonplace. A summary of the images present in this dataset is found in table 1 and figure 4, with more details below.

	2021 dataset	2025 dataset
Peak display luminance $[cd/m^2]$	$\approx 100 - 150$	1000
Min. luminance [% peak]	0.5 %	0.5 %
Max. luminance [% peak]	80 %	100 %
Number of patches N	93	102

Table 1. Description of the 2021 and 2025 datasets. The peak display luminance is indicative for previous experiments run on SDR displays, but exact for the 2025 dataset which is run on an HDR display. Other values are expressed in percentage of this peak.

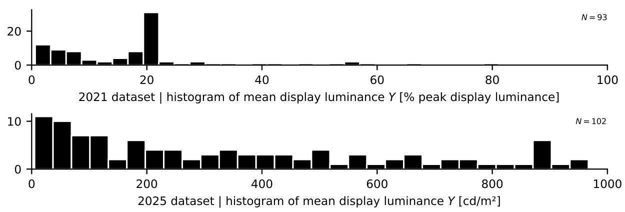

Figure 4. Histogram of patch mean luminances for both the 2021 and 2025 datasets.

At the end of this experiment, we want to answer the following questions using this data:

- do we need a new Visual Noise formula, or are the previous formulae good enough ?
- does the luminance of the scene (from the patch and background) influence perception, and by how much ?
- does a different colorspace give us more accurate results or simpler formulae than CIELAB ?

Image selection

At DXOMARK, several setups are used for the computation of Visual Noise: figure 5 shows two of them.

Figure 5. Autofocus HDR setup (left), Visual Noise chart (right)

From a selection of videos and photos of these setups recorded by several different smartphones in various lighting conditions, a total of 1602 patches (for the Visual Noise chart) and 3104 patches (for the Autofocus HDR setup) uniform gray patches are extracted and saved as lossless 16-bit PNG images with ICC profiles containing CICP metadata using the PQ (Perceptual Quantizer) EOTF and ITU-R BT.2020 colorimetry, which are supported by many HDR image viewers. These images are augmented and filtered as described in the following paragraphs to form the final dataset used during the experiment.

Patch size While the framing of these lab setups is always the same, extracted patches had different sizes because the devices recorded at different resolutions. We eliminated patches that were too small ($< 100 \times 100$ pixels), and extracted patches were rescaled to a common resolution in order to enable proper comparison. The target size was chosen to be 300 pixels, as it lies between the two original patch sizes (typical sizes are approximately 100 pixels for the Autofocus HDR setup and 450 pixels for the Visual Noise chart).

Patch luminance Because of size constraints, most of the charts these patches were extracted from are reflective objects, which are much darker than the maximum luminance found in the scene ; as a result, the maximum patch luminance available was around 400 nits, which was not sufficient to meet our objective.

To address this limitation, we augmented the dataset in several bins from 400 to 1000 nits, by multiplying each patch by a linear scaling factor based on the average patch luminance Y such that $Y_{\text{target}} = \text{factor} \cdot Y$, where the target luminance Y_{target} was randomly chosen within the desired range for each bin.

Gradient correction Given how significant the impact of non-uniformities is, we will correct gradients on 90% of the images using algorithm 1. The remaining 10% of the images are kept as-is so that the impact of gradient correction can be measured.

Final selection The selection of $N_0 = 100$ patches was carried out according to the following criteria:

- Remove patches with visible defects (e.g. strong lighting problems or device artifacts which were clearly not perceived as noise).
- Use patches where both `ISO2023` and `DXOMARK2021` provide consistent measurements (within ± 1 JND) and divide the space of 8 JNDs (according to `DXOMARK2021` JND values) into 10 equally sized bins to test all possible intensities of noise.
- Within each bin, randomly select a fixed number of patches using the average display luminance Y of each patch, mapped in such a way as to give more weight to the low end of the luminance curve.

2 additional patches with zero noise were added, with display luminances respectively 10 nits and 203 nits, intending to help anchor the experiment as well as the final formulas: a well-behaved Visual Noise metric is expected to be zero for them.

Experiment setup and annotations

Experimental procedure The main psychovisual experiment described in this paper takes the form of a 2 alternatives forced choice (2AFC) pairwise comparison. After a short introduction to the purpose of the experiment, each observer is presented a series of comparisons between two patches, and asked each time ”Choose the noisiest patch in each pair”.

Construction of a JOD scale This experiment was performed 2 times with different viewing conditions, described in table 2, each having more than 1 standard trial per experiment, which corresponds to all the possible comparisons for a N-image set, or $N(N-1)/2 = 5151$ pairs. A psychometric scale was constructed by statistically encoding all annotators' preferences, adhering to Thurstone's Case V model, as outlined in [9]. Paired comparison results are converted into Just Objectionable Difference (JOD) units, representing the average perceptual judgment across observers. A difference of 1 JOD corresponds to 75% of observers perceiving a difference, while a random guess (50% probability) equals 0 JOD.

ASAP Instead of doing random comparison between patches, we will use the *ASAP* algorithm [10], whose goal is to reduce the number of comparisons required to reliably converge to the final

Viewing distance	650 mm from the screen		
Display model	Apple Pro Display XDR		
Display profile	ISO 22028-5 reference mode		
Background (cd/m^2)	4.93	193	Measured using a Gossen Maxolux luminance-meter placed in front of the display
Bias light (lux)	5	23	Measured using a Gossel Mavolux illuminance-meter placed on the forehead of a participant
Observers	36	41	Experience ranging from image quality experts to novices
Comparisons	5849	5504	

Table 2. *Experiment setup and viewing conditions for the two phases ("dark" and "grey" background) of the pairwise comparison study.*

results. The algorithm adaptively selects the next pair of stimuli to be compared based on the current state of the model, focusing on those regions of uncertainty. This allows the method to efficiently explore the perceptual space while minimizing redundant comparisons.

Experimental results Each phase of the experiment resulted in an independent JOD scale represented in figure 6.

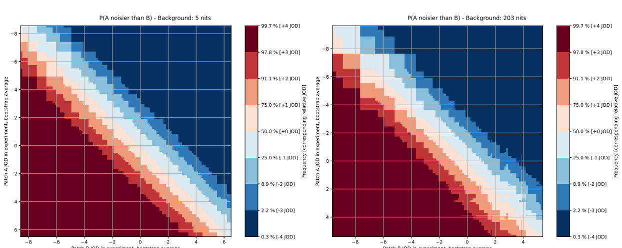

Figure 6. *Representation of the JOD rulers computed after (a) the "dark background" experiment, on the left and (b) the "grey background" experiment, on the right. For each figure, patches are placed non-uniformly on both the X and Y axis according to their JOD value in the corresponding phase of the experiment ; the percentage of observers choosing one patch over another is mapped to colors representing each frequency, or equivalently JOD level.*

Derivation of a Visual Noise metric

Formula exploration In the first part of the exploration, we output for each patch all available statistics that could be used: mean, variance, covariance, in addition to background luminance. These were performed in 4 color spaces (XYZ, CIELAB, AC_1C_2, IC_TC_P), for 3 high-pass filters (`CPIQ2023`, `DXOMARK2021` and `ISO2023`, which uses none), and with/without gradient correction ; each combination is fitted to the JOD ruler using a symbolic regression package, PySR [11], designed to automatically discover many mathematical expressions that describe a given dataset.

The main interesting features that came out in all models are luma variances (VarL^*, Var$A$, Var$I$ or Var$Y$) as well as averages. Background luminance, covariances and chroma variances were always much less significant (and probably the result of an over-fitting).

Gradient correction provides a more reliable explanation of the data in all models. Similarly, the CSF is present in most of the best-performing models. The choice of high-pass filter, however, does not appear to have a significant impact for this dataset, as all variants are represented equally.

It is interesting to see that, in spite of our initial hypothesis, the background luminance (which we hoped to relate to the adaptation state of the observer) seems to have very little effect on this dataset.

Formula selection From this exploration, simpler final formulae were selected and further optimized using the least-squares regression module from SciPy [12]. From these, we retain the first 3 formulae in table 3 as candidate `DXOMARK2025` Visual Noise metrics:

- (1), in AC_1C_2 space, is among the best-performing, taking into account variance and average of the patch display luminance A ;
- (2), in the IC_TC_P space, taking into account only the I channel, is particularly attractive for its simplicity ;
- (3) adds chroma variances to the same formula as (2) even though it performs slightly worse : a good metric should output a non-zero value for a patch with only chroma noise, which is the case for neither (1) nor (2).

	Formula
(1)	$\ln\left[\mathrm{Var}_A^2/\left(A^\beta+b\right)^2+e\right]+C$
(2)	$\ln\left[\mathrm{Var}_{I^*}^2+e\right]+C$
(3)	$\ln\left[\sum \mathrm{Var}_{I^*T^*P^*}^2+e\right]+C$

	(1)	(2)	(3)
b	124.3	-	-
β	1.765	-	-
e	2.05e−10	6.74e−4	7.30e−4
C	13.5	-1.37	-1.65

Table 3. *`DXOMARK2025` candidate model formulae, with the values of their parameters.* Var_A *is the variance of the* A *channel in* AC_1C_2 *color space (equal to* Y *in the* XYZ *color space) ;* A *is the average value of the same channel.* $\mathrm{Var}_{I^*} = \mathrm{Var}\left[720\cdot I\right]$ *is the variance of the* $I^* = 720\cdot I$ *channel in* IC_TC_P *colorspace, and* $\sum \mathrm{Var}_{I^*T^*P^*}^2 = \mathrm{Var}\left[720\cdot I\right]^2+\mathrm{Var}\left[360\cdot C_T\right]^2+\mathrm{Var}\left[720\cdot C_P\right]^2$.

Filtering method The impact of filtering algorithms is summarized in table 4. On the 2025 dataset, candidate models appear to perform similarly when gradient correction is enabled, and significantly worse when it is not. The model results will therefore be presented with the same filtering as `ISO2023` (i.e. only CSF, without any HPF) and gradient correction enabled.

Non-uniform (and therefore non-periodic) patch edges also influence the application of filters in Fourier space by producing unwanted "oscillations" on the entire patch. This remains a problem that is not fully mitigated by the various high-pass filters in their original publications. We attempted mitigation strategies based on windowing functions, but these were not found significant enough when gradient correction is enabled to be included in these results.

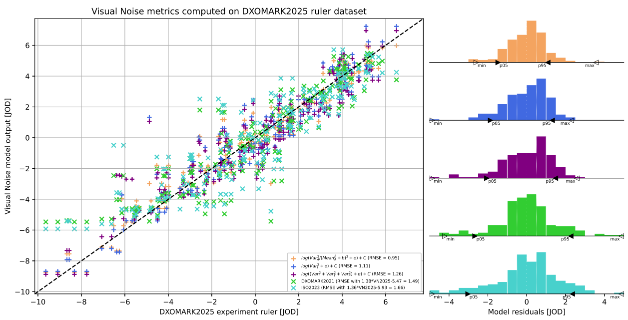

Figure 7. A comparison between candidate Visual Noise 2025 metrics, `DXOMARK2021` and `ISO2023` metrics, on the 2025 JOD ruler. The figure shows (a) on the left, a scatter plot between the JOD output of the corresponding psychovisual experiment and the metric predicted by the formula ; (b) on the right, the distribution of residual errors between the "ground-truth" JOD and the value predicted by each candidate formula. For existing metrics, an affine mapping is determined to match the 2025 JOD scale, whose coefficients and RMSE are reported in the legend.

			`DXOMARK2025` candidate model		
CSF	**HPF**	**Correction**	**(1)**	**(2)**	**(3)**
ISO	DXOMARK [1]	Yes	0.92	1.10	1.29
CPIQ	CPIQ [2]	Yes	0.93	1.10	1.15
ISO	ISO [3]	Yes	0.95	1.11	1.26
ISO	ISO [3]	No	1.31	1.41	1.26

Table 4. Performance of `DXOMARK2025` candidate model formulae with using the same filtering methods as CPIQ Visual Noise [2], `ISO2023` [3] and `DXOMARK2021` [1].

Comparison with the already available Visual Noise metrics Figures 7 and 8 show a comparison between the existing Visual Noise metrics on the 2021 and 2025 datasets. To compute existing metrics based on the CIELAB space on HDR images, a fixed white point luminance of 1000 nits was used.

Our best new candidate Visual Noise metric, `DXOMARK2025` (1), has similar performance as `DXOMARK2021` and `ISO2023` for ranking patches on the 2021 dataset even though it wasn't trained on it. It performs significantly better than both of them on the 2025 dataset, especially for patches with low amounts of noise.

`DXOMARK2025` (2) has similar performance as `ISO2023` on the 2021 dataset and acceptable performance on the 2025 dataset.

Finally, and contrary to one of our initial hypotheses, with the exception of a few outliers in the 0-5 nits range or close to the JND which will require further study, adaptation luminance does not seem to be significant for all other patches in our dataset.

Note on JOD and JND scales Unfortunately, in these comparisons, the JOD scale resulting from the 2025 experiments only matches the 2021 JND scale up to affine transformations, the alignment of which will be the topic of a future research.

Ideally, to transform the ruler from a relative JOD scale (centered around 0) to a scale anchored in perceptual units (JND), an additional experiment is required to precisely locate the JND threshold. This step is crucial because the mapping function that converts choice probabilities to JOD distances becomes unstable at extreme values, which can lead to overestimation of gaps. The experiment of converting to JND stabilizes distances at the ex-

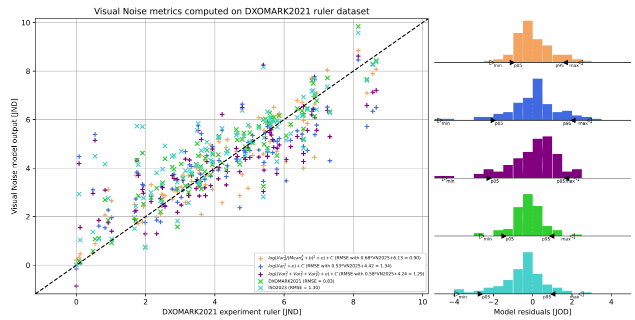

Figure 8. Similar comparison as figure 7, comparing the same metrics on the 2021 JND ruler. For 2025 metrics, an affine mapping is determined to match the 2021 JND scale, whose coefficients and RMSE are reported in the legend.

tremes and anchors the scale to a reliable perceptual threshold.

The results of our experiments also having a significant *scale* difference compared to previous metrics hint that a hidden parameter of the experiment is not taken into account by the model – be it the viewing conditions or the way patches are presented ; for example, previous psychovisual experiments used the ISO triplet comparison [13] method, unlike this one. It is however interesting to note that only an affine relationship is required to match scales: noisiness ordering and approximate perceptual linearity is preserved.

Conclusion

We think a new Visual Noise formula will help the evaluation of current camera systems with HDR capabilities: a better gradient correction is necessary in any case to handle images captured in the wild by modern processing pipelines on scenes with higher dynamic range, and a promising candidate metric based on simple luminance statistics outperformed previous CIELAB-based metrics for ranking the noisiness of patches while being fully compatible with both HDR and SDR media.

The luminance of the patch does influence the perceived noisiness to a large extent ; however, the luminance of the background has a much less visible influence in our dataset, and we cannot derive a conclusion on the impact of observer adaptation from these experiments alone.

Future work

Several directions remain to be explored in order to strengthen and validate the proposed approach.

The derived formulae do not align to unitary JND scales used in existing Visual Noise formulations yet. Early attempts were made to derive a JND scale out of the JOD results from the 2025 experiments, but so far this only yielded poorer accuracy in the general case. Future research will determine an affine JND mapping that matches existing Visual Noise metrics, or find reasons for a potential discrepancy between the JOD scales.

The available datasets were also not able to produce conclusive evidence of the impact of observer adaptation on the luminance range they cover, contrary to what we hoped. A further study on this topic will require construction of an extended dataset with a larger patch and background luminance range, and compare the results with other psychovisual studies and models.

Finally, chroma variance is obviously an important aspect in a Visual Noise model (a patch with pure chroma noise and perfectly uniform luma should return a non-zero Visual Noise metric), but the available datasets were not able to produce conclusive evidence either. Future studies will incorporate a larger selection of chroma noise in the dataset in order to attempt to model this.

Algorithm 1 Algorithm used for gradient estimation and non-uniformity correction in `DXOMARK2025` metrics. To perform this correction, we fit a 2D plane on the Y channel in the linear $\mathtt{XYZ}_{D65}$ color space, invert it, and apply the correction to the image. See example output in figure 9.

function CORRECTGRADIENT(image)
  $(H,W,C) \leftarrow \text{dimensions}(image)$
  $Y \leftarrow image[:,:,1]$ ▷ Use Y channel
  $x \leftarrow \text{linspace}(-0.5, 0.5, W)$
  $y \leftarrow \text{linspace}(-0.5, 0.5, H)$
  $(xv, yv) \leftarrow \text{meshgrid}(x, y)$
  $X \leftarrow [xv.\text{flatten}, yv.\text{flatten}, \text{ones}(H \times W)]$
  $y\_vals \leftarrow Y.\text{flatten}$
  $\text{coef} \leftarrow \text{LeastSquares}(X, y\_vals)$
  $plane \leftarrow \text{coef}[0] \cdot xv + \text{coef}[1] \cdot yv + \text{coef}[2]$
  $mean\_plane \leftarrow \text{mean}(plane)$
  $gradient \leftarrow plane / mean\_plane$
  $corrected \leftarrow image / gradient[:,:,\text{newaxis}]$
  return corrected
end function

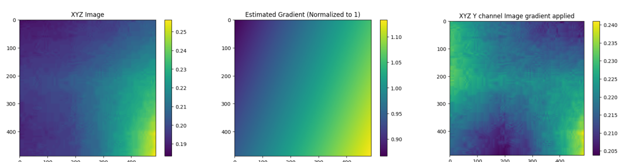

Figure 9. *Example input image presenting non-uniformities (left), estimated gradient map (middle), and corrected image (right).*

Contributions

Hugo Masson : Conceptualization, Methodology, Investigation, Software, Formal analysis, Writing – original draft.

François-Xavier Thomas : Conceptualization; Supervision, Methodology, Validation, Writing – review & editing.

Claudio Greco : Validation, Formal analysis, Writing – review & editing.

Daniela Carfora Ventura : Methodology (pairwise comparison), Software (ASAP code), Data curation, Formal analysis.

Mauro Patti : Resources, Validation (HDR-related aspects).

Other supervisors : Supervision, Project administration.

Bibliography

[1] Thomas Bourbon et al. "New visual noise measurement on a versatile laboratory setup in HDR conditions for smartphone camera testing". In: *Electronic Imaging 2022, Image Quality and System Performance XIX*. Burlingame, CA, USA: Society for Imaging Science and Technology (IS&T), 2022, pp. 52-1–52-8. URL: https://corp.dxomark.com/wp-content/uploads/2022/02/EI2022_VisualNoise_Article.pdf.

[2] IEEE Standards Association. *IEEE Standard for Camera Phone Image Quality (CPIQ)*. Aug. 2023. URL: https://standards.ieee.org/ieee/1858/6931/.

[3] International Organization for Standardization (ISO). *Photography — Electronic still-picture imaging — Noise measurements*. Apr. 2023. URL: https://www.iso.org/standard/82233.html.

[4] International Organization for Standardization (ISO). *Photography and graphic technology — Extended colour encodings for digital image storage, manipulation and interchange — Part 5: High dynamic range and wide colour gamut encoding for still images (HDR/WCG)*. June 2023. URL: https://www.iso.org/standard/81863.html.

[5] *The CIE 2016 Colour Appearance Model for Colour Management Systems: CIECAM16*. 2022. DOI: 10.25039/tr.248.2022. URL: https://cie.co.at/publications/cie-2016-colour-appearance-model-colour-management-systems-ciecam16.

[6] Rafal K. Mantiuk, Dounia Hammou, and Param Hanji. *HDR-VDP-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content*. 2023. URL: https://arxiv.org/abs/2304.13625.

[7] Peter Vangorp et al. "A Model of Local Adaptation". In: *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2015)* (2015). URL: https://doi.org/10.1145/2816795.2818086.

[8] Pontus Andersson et al. "ꟻLIP: A Difference Evaluator for Alternating Images". In: *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 3.2 (2020), 15:1–15:23.

[9] Maria Perez-Ortiz and Rafal K. Mantiuk. *A practical guide and software for analysing pairwise comparison experiments*. 2017. DOI: 10.48550/ARXIV.1712.03686.

[10] Aliaksei Mikhailiuk et al. *Active Sampling for Pairwise Comparisons via Approximate Message Passing and Information Gain Maximization*. 2020. arXiv: 2004.05691 [cs.LG]. URL: https://arxiv.org/abs/2004.05691.

[11] Miles Cranmer. "Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl". In: (May 2023). DOI: 10.48550/ARXIV.2305.01582. arXiv: 2305.01582 [astro-ph.IM].

[12] Pauli Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature Methods* 17.3 (Feb. 2020), pp. 261–272. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2.

[13] International Organization for Standardization (ISO). *Photography — Psychophysical experimental methods for estimating image quality — Part 2: Triplet comparison method*. 2005. URL: https://www.iso.org/standard/38332.html.