# Image Quality Assessment for Natural Scene Portraits: An Industrial Application

***Daniela Carfora Ventura, Gabriel Pacianotto Gouveia, Hoang-Son Nguyen, Jianqiang Sky Zhou, Nicolas Chahine, Sira Ferradans, DXOMARK Image Labs, Paris, France***

## Abstract

*Portraits are one of the most common use cases in photography, especially in smartphone photography. However, evaluating portrait quality in real portraits is costly and difficult to reproduce. We propose a new method to evaluate a large range of detail preservation rendering on real portrait images.*

*Our approach is based on 1) annotating a set of portrait images grouped by semantic content using pairwise comparison 2) taking advantage of the fact that we are focusing on portraits, using cross-content annotations to align the quality scales 3) training a machine learning model on the global quality scale.*

*On top of providing a fine-grained wide range detail preservation quality output, numerical experiments show that the proposed method correlates highly with the perceptual evaluation of image quality experts.*

Keywords: Image quality evaluation, Texture detail preservation, Data set fitting, DXOMARK, Machine Learning, Computer Vision, Realistic image quality, Camera quality assessment

## Introduction

Camera quality is a major reason for consumers to choose between smartphones. However, evaluating a camera's quality is costly, cumbersome, and, in many cases, a non-repeatable process. To simplify and improve this process, many techniques have been developed for automatic image quality assessment.

In the state of the art, image detail quality assessment is predominantly based on objective metrics applied to synthetic patterns, such as the widely used Dead Leaves chart, which simulates real-world characteristics [3]. While these methods are well-established, they struggle to keep pace with the growing complexity of modern imaging systems. As camera technologies evolve, integrating sophisticated algorithms that dynamically adapt to scene content, the relevance of traditional evaluation techniques comes into question. These new computational pipelines demand more nuanced approaches to accurately measure the enhancements offered [5].

Evaluating portrait rendering using face-centered content in the laboratory, such as the usage of mannequins [4,7], marks a clear advancement over conventional flat chart-based techniques. The application of machine learning enables a deeper investigation into camera optimizations designed specifically for facial imagery [4]. By integrating face-specific content with machine learning algorithms, assessments can more accurately reflect the capabilities of modern camera systems, considering factors like noise management and image sharpening [4,7].

However, laboratory setups cannot reproduce the complexities of real-life scenes. This study continues extending the previous works [2,6] that bridge the gap from laboratory metrics [4,7] to real scenes for portrait assessment using machine learning models.

**Contributions**: To ensure high-quality ground truth labels, we introduce a structured shooting protocol and evaluation protocol. The shooting protocol defines a series of scenes characterized by their content, framing, illumination conditions, and camera parametrization. The evaluation protocol specifies the visualization conditions, the annotation task – in this case, a pairwise comparison - and the analysis conducted by image quality experts based on the question posed.
This protocol ensures repeatability and precision in the image annotation, as experts compare the same content and answer to the same feature analysis. However, this process creates scene-dependent quality scales that are not aligned across scenes even though face crops are scaled to be viewed at the same size. Such misaligned ground truth poses significant challenges for training machine learning models.
To address this, we propose to construct a global portrait scale for detail rendition by integrating per-scene scales through cross-content analysis. This method preserves the granularity of per-scene pairwise comparisons while enabling the creation of a common perceptual quality scale suitable for training machine learning models. Moreover, this global scale not only facilitates more effective model training but also ensures fair comparisons between scenes in both perceptual and scoring terms, a critical requirement for industrial applications where consistency across diverse content is essential.

## Novelty

The study expands upon previous work on detail preservation assessment in real-scene portraits [6,2]. In this paper, we propose a new approach to deal with the variety of scenes that compose the PIQ data set[6]. The proposed method allows us to maintain the precision and granularity of the original scenes, while also having a global perspective in the quality scale. This is particularly valuable for industrial applications where fair, consistent, and accurate cross-scene comparisons are essential. The numerical results show a strong correlation with human perception on unseen scenes, which indicates that the current approach allows generalization outside the training scenes. This ability to extend across diverse portrait scenes reinforces the robustness of the model for real-world deployment.

## Proposed method

The publicly available dataset proposed in [6] provides for every image a quality score measured in Just Objectionable Difference (JOD) units [1]. These scores are obtained from pairwise comparisons conducted by more than 30 experts. In this study, we

will focus on the detail preservation feature (texture-noise compromise) and 25 selected scenes.

**Database extension:** To extend the quality range, we selected the 25 most diverse and relevant scenes from PIQ, ensuring a broad representation of conditions for industrial case interests. The goal of this extension is to enhance the granularity of the entire dataset and extend the quality range at the high end of the scale.
To this end, we:

1) Increased the diversity of devices by adding several smartphones to the dataset.
2) Introduced high-quality references from a DSLR (Sony A7R4).
3) Generated synthetic images to bridge the gaps in the quality scale. We use subsampling, slight sharpening, and Gaussian noise. This simple approach ensured a smooth transition between high-quality DSLR images and smartphones, using fixed variance noise for controlled degradation of texture-noise compromise.

As a result, the total number of images per scene has approximately increased from 100 to 200.

**Annotations:** Our annotation process follows the same controlled viewing conditions outlined in PIQ23 [6], to ensure perception-based quality annotations under consistent and unbiased conditions.

- The new images per scene were integrated into the previously annotated dataset in the per-scene annotation task. These new annotations allow us to update the JOD scene scale for each of the 25 scenes.

- Moreover, we sampled images from each scene along the quality scale to create a cross-scene (cross-content) image set. This set was annotated under the same conditions to maintain consistency.

**Alignment of the scene-scales:** Every image $i$ annotated on the scene-annotation task has a JOD score $x_i$. From the cross-scene annotation task, we also obtain a set of JOD scores for every image ($y_i$). Therefore, this set of images has two scores ($x_i$ and $y_i$) that we can use to align the scales. For every scene $j$, we select the images with both annotations and compute the alignment between them using least squares:

$$y_{i,j} = a_j x_{i,j} + b_j \quad (1)$$

**Creation of a unified scale:** In order to create the common unified scale, we use the previously calculated coefficients $a_j$, $b_j$. It is important to notice that the cross-content scale $y$ was created from images that did not share the same framing or illumination condition, and therefore are more difficult to annotate. This means that the final scores $y$ are less precise and granular than those obtained from the per-scene annotations $x$. Therefore, aligning directly with the cross-content scale $y$ is not ideal. Instead, it is crucial to identify a reference scene $x_j$. This reference scene should ideally have a wide JOD range, as it relates to various image renditions and diverse skin tones. Once this scene is selected, we can obtain the aligned ground truth scores by solving the associated linear equation. In Figure 1 we can observe a sample of the produced quality scale. For space reasons, the images needed to be reduced. To better observe the difference between the images in the high-end region please refer to: https://corp.dxomark.com/image-sample-from-the-psychophysical-scale/.

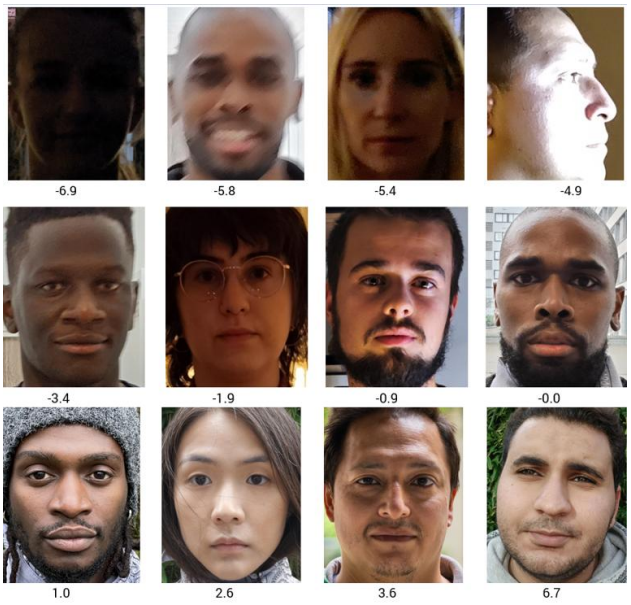


***Figure 1**. Sample of images from the aligned scale and their associated JOD detail quality score. Note that 1 JOD distance implies that the difference in quality between the images is visible. For high-resolution copies of the images, please refer to this link*

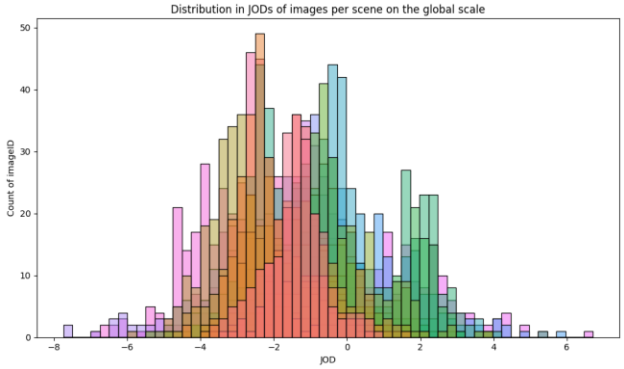


***Figure 2**. Histogram along the JOD quality scale of the aligned scenes. The color indicates a scene.*

**Data set split:** A main objection of using scene-dependent scales is the fact that given a new scene, the ML model needs to be retrained on the new scene so that it can learn the associated scale. However, the characteristics that make an image better or worse in terms of image quality do not depend on the scene. To evaluate the capacity of the ML model to generalize to new scenes, we propose a train-valid-test scene split based on scenes. Inspired by [2], this means that the scenes in the training set are not on the testing set and vice versa.

**Model training:** To assess the impact of scene-dependent scales and the benefits of the alignment, we trained multiple ML architectures using both the individual scene scales and the aligned global scale on the scene-split test set.

We employed a multitask approach inspired by FULL-HyperIQA [10] for the individual scene scales, enabling the model to learn the relationship between scale and scene. However, given the challenge of classifying the scene solely from a cropped face region, we modified the architecture to incorporate the complete image as additional input (see Figure 3). This variant is referred to as the FULL-HyperIQA variant in Table 1.

The loss function combined Mean Squared Error for quality prediction and Cross-Entropy loss for scene classification. We applied a weighted sum, assigning a weight of 0.5 to the classification and 1.0 to the quality prediction loss.

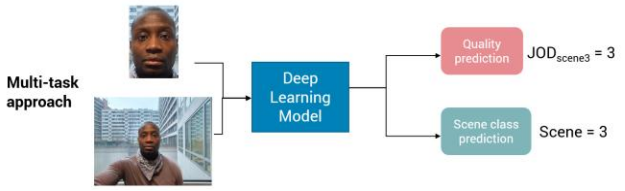


***Figure 3.*** *Multitask approach to train on the individual scene scales.*

By contrast, training on the aligned global scale simplifies the task, as the model only needs to predict a single quality score without learning scene dependencies. In this case, we used a single-task model optimized purely for quality prediction (see Figure 4).


***Figure 4.*** *Single-task approach to train on the unified scale.*

To ensure robust evaluation, we tested different training configurations for all the proposed methods and reported the best results. We randomly cropped square patches of size 1344 (1 patch per image) and used Adam stochastic optimization with different learning rates between $10^{-6}$ and $10^{-4}$. The training was conducted for 300 epochs with a learning rate decay factor of 0.05 every 10 epochs. For FULL-HyperIQA, we experimented with different numbers of scenes (values of k) including 3, 5, 10, and 25. Early stopping was applied with a patience of 40 epochs.

## Results

The new quality scale extends 14 JOD units, a wide JOD range that goes from low-light conditions to outdoor well-illuminated scenes (See Figures 1 and 2).

The alignment between the original per-scene scales and the cross-content scale is illustrated in Figure 5, where the scaling coefficients *(a, b)* are plotted for each scene. Ideally, a perfect alignment would result in *a=1* and *b=0*. The deviation of these coefficients from this perfect fit highlights the differences in scene-specific quality scales, underscoring the importance of alignment.

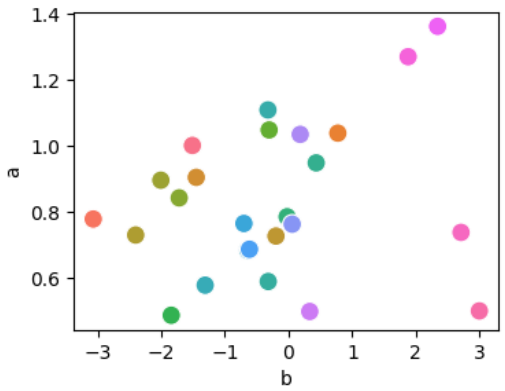


***Figure 5****. Representation of the a,b coefficients that align scales (see text for more details). Color represents different scenes.*

Table 1, compares models trained on individual scene scales versus those trained on the global aligned scale, on the same testing set. To ensure a fair comparison, we compute metrics for each scene separately and report the average performance across all scenes. The results demonstrate that the models trained on the unified scale generalize well to unseen scenes, showing a strong correlation with human perceptual evaluations (see SRCC column).

Furthermore, Table 2 provides a detailed breakdown of the unified scale, with metrics computed globally (without scene notion). Notably, the mean absolute error (MAE) between the predicted values and the ground truth is, on average, below 1 JOD for MUSIQ and HyperIQA backbones. This means that this error is, in general, not perceivable.

| Data | Model | PLCC | SROCC |
|---|---|---|---|
| Individual scene scales | Multitask resnet50 baseline | 0.60 | 0.62 |
| | FULL-HyperIQA [10] | 0.63 | 0.67 |
| | FULL-HyperIQA variant | 0.71 | 0.72 |
| Unified scale | Resnet50 baseline | 0.78 | 0.76 |
| | MUSIQ [9] | 0.82 | 0.78 |
| | HyperIQA [8] | **0.83** | **0.80** |

***Table 1.*** *Comparison of baselines according to their average scene correlation (PLCC and SROCC). As shown by the table, the deep learning methods tested perform significantly better on the unified scale.*

| Data | Model | SROCC | MAE |
|---|---|---|---|
| Unified scale | Resnet50 baseline | 0.81 | 1.41 |
| | MUSIQ [9] | **0.90** | **0.47** |
| | HyperIQA [8] | 0.88 | 0.58 |

***Table 2.*** *Performance of deep learning models trained on the aligned scale. Results are computed on the scene-split test set (see text for more details)*

Overall, the results in Tables 1 and 2 highlight the benefits of using the unified scale. This approach not only improves generalization to unseen scenes but also simplifies the training process.

## Conclusions

This work extends an existing portrait dataset by incorporating both real and synthetic portraits, annotating portrait scenes, constructing an extended JOD-based quality scale, and training ML architectures for image quality prediction. We introduced a methodology for:

- Constructing a psychophysical quality scale for portrait images,
- Evaluating the detail-noise tradeoff independently of scene variations.

Numerical results demonstrate:

- The effectiveness of the proposed method in modeling perceptual quality.
- The superiority of scene alignment over multitask learning when dealing with different JOD scene scales.

The proposed methodology represents a significant contribution to the field of portrait image quality assessment, offering a robust framework for evaluating image quality across diverse conditions.

## References


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

[2] Chahine, N., Conde, M.V., Carfora, D., Pacianotto, G., Pochon, B., Ferradans, S., Timofte, R., Duan, Z., Xu, X., Huang, Y. and Yuan, Q., 2024. Deep portrait quality assessment. a NTIRE 2024 challenge survey. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6732-6744).

[3] Cao, Frédéric, Frédéric Guichard, and Hervé Hornung. Measuring texture quality of a digital camera. In Digital Photography V, vol. 7250, p. 72500H. International Society for Optics and Photonics, 2009.

[4] C Nicolas, B Salim. “Portrait Quality Assessment using Multi-Scale CNN”. In London Imaging Meeting (Vol. 2021, No. 1, pp. 5-10). Society for Imaging Science and Technology

[5] Van Zwanenberg, Oliver, Sophie Triantaphillidou, Robin Jenkin, and Alexandra Psarrou. “Edge detection techniques for quantifying spatial imaging system performance and image quality”, (2019).

[6] Chahine, N., Calarasanu, S., Garcia-Civiero, D., Cayla, T., Ferradans, S., and Ponce, J. “An Image Quality Assessment Dataset for Portraits”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2023).

[7] Ventura, Daniela Carfora, Gabriel Pacianotto Gouveia, Ana Calarasanu, Valentine Tosel, Nicolas Chahine, and Sira Ferradans. "From Video Conferences to DSLRs: An In-depth Texture Evaluation with Realistic Mannequins." Electronic Imaging 36 (2024): 1-6.

[8] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3667±3676, 2020. 1, 3, 8

[9] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5148±5157, 2021. 1, 3, 8

[10] Nicolas Chahine, Sira Ferradans, Javier Vazquez-Corral, Jean Ponce. Generalized Portrait Quality Assessment. 2024. ffhal-04457073


## Acknowledgments


This project was provided with computer and storage resources by GENCI at IDRIS thanks to the grant 2024-AD011014305R1 on the supercomputer Jean Zay's the V100.


## Author Biography

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

***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 has worked as a Machine Learning Engineer, helping to develop new image assessment techniques with the use of machine learning.*

***Hoang-Son Nguyen*** *holds a master's degree from Sorbonne University in data, knowledge, and computer science. Intern at DXOMARK during 2024.*

***Jianqiang Sky Zhou*** *earned his Ph.D. in condensed matter physics from École Polytechnique in 2016, after which he continued his research as a postdoctoral researcher. Since 2019, he has been working as a machine learning engineer specializing in computer vision. In 2024, he joined DXOMARK as a machine learning engineer, applying his expertise to image quality assessment*

***Nicolas Chahine*** *is a machine learning Ph.D. student. He followed a double degree program between the Lebanese university faculty of engineering and Telecom Paris (2014-2020). He also followed a master’s degree in applied mathematics, namely MVA, at the University of Paris Saclay in collaboration with Ecole Normale Superieur (2019-2020). In November 2024, he defended his thesis on Portrait Image Quality Assessment, developed in collaboration with DXOMARK and INRIA.*

***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.*