From Video Conferences to DSLRs: An In-depth Texture Evaluation with Realistic Mannequins

Daniela Carfora Ventura, Gabriel Pacianotto Gouveia, Ana Calarasanu, Valentine Tosel, 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, inconvenient, and difficult to reproduce. We propose a new method to evaluate a large range of detail preservation renditions on realistic mannequins. This laboratory setup can cover all commercial cameras from videoconference to high-end DSLRs. Our method is based on 1) the training of a machine learning method on a perceptual scale target 2) the usage of two different regions of interest per mannequin depending on the quality of the input portrait image 3) the merge of the two quality scales to produce the final wide range scale.
On top of providing a fine-grained wide range detail preservation quality output, numerical experiments show that the proposed method is robust to noise and sharpening, unlike other commonly used methods such as the texture acutance on the Dead Leaves chart.

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

Introduction

In the ever-evolving field of photography, portraits have become a prominent focus of interest, especially with the rise of smartphone photography. Traditionally, objective evaluation of detail preservation has mainly centered on non-statistical methods. These methods often use synthetic content, such as edges or patterns that can simulate, in some cases, properties of natural images, as seen in the Dead Leaves chart [3][5].

When dealing with synthetically generated visual charts, captured in controlled laboratory conditions, measures such as Noise Power Spectrum (NPS) and the Modulation Transfer Function (MTF), have been used to evaluate image quality attributes (e.g., noise, texture, and sharpness). Studies done on raw images captured with DSLRs, in manual mode [1][3] show satisfactory results. However, these measures correlate relatively poorly with human perception when including nonlinear camera processes such as multi-image fusion or deep learning-based image enhancement [6], which are widely known to be used by today's cameras.

The acutance, which combines the Modulation Transfer Function (MTF) and the contrast sensitivity function (CSF), is known to reflect the rendering of the quality of the photographic device. Early methods use charts containing a blur spot or a slanted edge to compute the MTF. Cao et al. [3] using the Dead Leaves chart, propose a more appropriate method for describing fine detail rendering. The MTF is then computed using the ratio of the reference chart and the respective Power spectral density (PSD) of the photographed image. This MTF is referred to as "Texture MTF" since it was computed from a textured area. However, this method assumes a linear optical system model. This is no longer a plausible assumption given the highly nonlinear processing done by current smartphone cameras, by, for instance doing sophisticated noise reduction processing.

Moreover, current smartphone's ISPs (Image processing pipelines) adapt to the content of the scene, especially when it involves people. Consequently, relying solely on synthetic visual content to assess camera devices, is not sufficient to capture the intricate behavior of modern imaging systems.

FIGURE 1 REALISTIC MANNEQUINS COVERED ON THE PROPOSED MODEL

To address these limitations, realistic mannequin setups (see Figure 1) were proposed [1,4] to allow a closer to natural scene content that would activate the portrait-mode, that is, the complex behavior of cameras when treating faces. This approach not only brings us closer to real-world content but also provides a robust foundation for automating image quality assessment through the application of Machine Learning (ML) methods.

Contributions: This paper proposes a new method for extending the detail preservation image quality scale of a database considering two main ingredients: 1) The target region of interest of a scene and 2) an algorithmic pipeline for merging independent scores. We train a ML model for evaluating texture on three different realistic mannequins (see Figure 1), and two different regions of interest per mannequin. The produced final scale covers cameras from low quality video conference up to high end DSLRs. Moreover, we show that our method, unlike other texture measures such as texture acutance on the Dead Leaves chart, is robust to noise and sharpening.

Novelty

The study expands upon previous work on detail preservation assessment in realistic mannequin setups [4]. It extends the range of image quality by introducing the separation of low-quality (LQ) and

high-quality (HQ) datasets. This approach was inspired by the method proposed to evaluate the perceptual noise in this setup [1]. Additionally, we introduce an innovative and robust pipeline for merging these two datasets, each of which focuses on different regions of interest (ROI) in the image. Compared to previous works on texture evaluation using machine learning [4], the proposed pipeline proves to be more robust in terms of correlation.

Having a single model, and scale, that covers a large range of cameras, and a large range of framing, from video conference to DSLRs, specifically for face content, is something that has not been addressed before. Moreover, the idea of merging datasets can be extrapolated to other perceptual evaluation attributes, where different areas of interest or perceptual annotations are required.

Moreover, we show the pertinence of the measure when compared to other standard texture measures such as the acutance on the Dead Leaves chart. The proposed method shows robust results to degradations due to the addition of noise or sharpening, unlike the acutance.

Proposed method

Current public consumer cameras extend from very low-resolution low-quality videoconference devices up to high-end DSLR cameras. To compare detail preservation rendering in a way that is robust to changes in the object scale, we need to define a common meaningful region of interest (ROI) with a predefined size. Given the difference in empirical resolution, this can be challenging. Small crops in high-quality cameras provide granularity to distinguish details for high-end devices but are not meaningful for low-quality cameras. On the other hand, big ROIs which give granularity for low-quality devices, do not give enough resolution to distinguish higher-end cameras. To overcome these issues, we propose the following pipeline:

Definition of ROIs: The areas with more details within the image are selected for high-quality assessment. The beard and the eyebrows are the regions of interest (ROI) for male and female mannequins, respectively. As for low-quality evaluation, the entire face serves as the ROI for every mannequin. (see Figure 2).

FIGURE 2. TWO ROIS WERE DEFINED ON EACH MANNEQUIN TO CAPTURE DIFFERENT DETAIL LEVELS (A) ROI FOR LOW-QUALITY CAMERAS AND (B) ROI FOR HIGH-END CAMERAS FOR FINE DETAIL EVALUATION.

Image dataset construction: The database comprises images from 3 different mannequins (An old white male, a young Asian woman, and a deep-skinned woman, see Figure 1), annotated under two different conditions. With 1869 images, sourced from over 200 devices, we guarantee a comprehensive representation of the market's quality range. To enhance dataset diversity, we include shots with multiple framings, various lighting conditions (low light, indoor, outdoor), and camera types (main, selfie, video cam), covering both photo and video formats.

Perceptual annotations and JOD quality scale construction: To obtain a ground truth for the dataset, perceptual annotations are conducted with a predefined specific question under a pairwise mode, as illustrated in Figure 3. The annotation task was done under controlled conditions: 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 image was viewed with a cutoff frequency of 30 cycles per degree, at a distance to the eye of 65cm, on a 32''16:9, UHD 4K screen, thus, with a pixel pitch of 0.185.

The annotation task was done by 20 different people, with up to 1.5 standard trials per scene, where 1 standard trail equals all the possible comparisons for a n-image set: $n(n-1)/2$ pairs.

After finishing the pairwise annotations, the psychometric scale is constructed by statistically encoding all annotators' preferences, adhering to Thurstone's Case V model, as outlined in Perez-Ortiz et al.'s research [2]. In Thurstone's model, each image score is considered a normal random variable with a mean and standard deviation, representing the true mean quality score.

FIGURE 3. EXAMPLE OF A PAIRWISE COMPARISON TASK DONE BY ANNOTATORS.

The results of paired comparisons are scaled into *Just Objectionable difference (JOD)*. The final continuous scale represents the average opinions across multiple observers when they choose based on "which one is closer to the perfect quality reference?". Two stimuli are 1 JOD apart if 75% of observers can see the difference between them, and a random guess (i.e., probability of 50%) results in a JOD distance of 0 between the images.
Note that scaling method involves using the inverse Gaussian cumulative distribution to map probabilities to distances on the JOD scale. This mapping is unstable for high distances, therefore the JOD scale is most meaningful for distances below 2 JOD.

The pairwise comparison task is done on images on the same content, therefore each ROI has its own scale. The scales are shift invariant. Without loss of generality, we center all scales at 0.

Machine Learning model for quality prediction: To address the difference in image sizes and quality scales between the datasets, we implement a multitask learning approach. In this approach, the prediction of the quality score is supported by the classification of the dataset type as an auxiliary task. The combination of the two tasks allows for a more comprehensive model capable of generalizing across different mannequin types, shooting conditions, and ROIs.

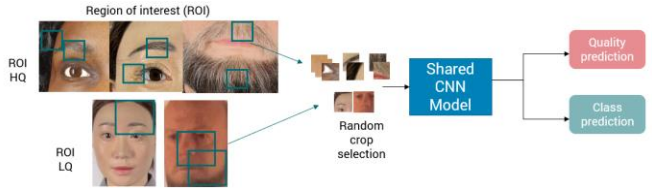

FIGURE 4. MULTITASK LEARNING APPROACH FOR QUALITY PREDICTION AND SCENE CLASSIFICATION

As can be seen in Figure 4, 30 random patches are taken from the input ROI and passed through the shared CNN model. We use the backbone of a pretrained ResNet-18 [7] to extract relevant features from the image. These extracted features from the input crop are then channeled into two separate fully connected heads – one for quality prediction and the other for class prediction. Each comprises three layers, ultimately resulting in the prediction of a float value for the texture quality and an integer value for the mannequin type.

Implementation details. The model is trained with a frozen backbone during the first 10 epochs and then unfrozen for the remaining training process. We employ Adam optimizer, with an initial learning rate of 10^{-4}. The loss is a sum of the Mean Squared Error (MSE) and Cross-entropy losses, ensuring proficiency in both tasks.

Creation of a unified scale: Each mannequin has two scales centered at zero, one for each viewing condition. We would like to compute the shift factor between the two scales to have a single continuous scale.

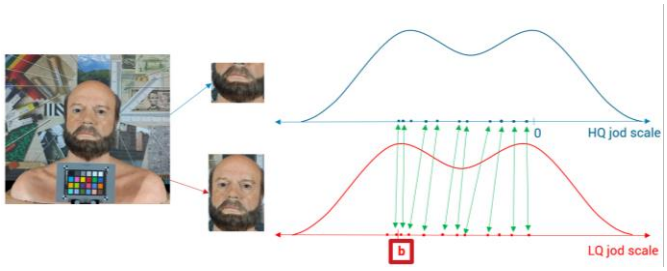

FIGURE 5. LQ AND HQ JOD SCALE ALIGNMENT PROCESS

Given a set of images of each mannequin, we crop the two ROIs, one for the Low Quality (LQ) scale and another for the High Quality (HQ) scale (see Figure 5 for an example). On each crop, we apply the trained model from the previous section. By averaging the differences of these output values for all the images, we can estimate the shift factor "b" that aligns the two scales.

Overall scale construction: These aligned scales yield comparable results, but each one of them is more precise in a region of the unified range. To avoid border effects, we propose an aggregated score computed as a smooth weighted average between low quality and high quality:

$$s_{aggr} = \frac{(s_{LQ} * w_{LQ} + s_{HQ} * w_{HQ})}{(w_{LQ} + w_{HQ})} \quad (1)$$

where s_{LQ} and s_{HQ} are the measurements of the HQ ROI and LQ ROI, respectively. The weight for HQ w_{HQ} increases with a higher HQ score, while the weight for LQ w_{LQ}, decreases proportionally, leading to a weighted average, as shown in Figure 6.

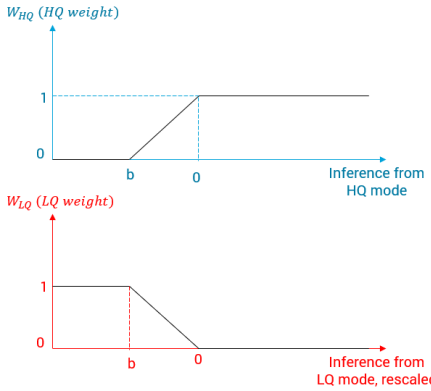

FIGURE 6 WEIGHTS DISTRIBUTION FOR THE AGGREGATED SCORE COMPUTATION

Results

Our Deep Learning (DL) model shows a strong correlation with human perceptual evaluations. Table 1 presents the linear and rank correlations (LCC and SROCC), together with the Mean Absolute Error in JOD (MAE). The multitask training strategy gives a high performance individually on each dataset. This enables us to later merge their scales accurately. A very small MAE implies most points are inside our interval [-0.5, 0.5], below which a person (on average) cannot distinguish a change in quality.

Dataset	Regression metrics		
	LCC	SROCC	MAE (JOD)
High quality (HQ)	0.978 *(± 0.005)*	0.975 *(± 0.001)*	0.318 *(± 0.001)*
Low quality (LQ)	0.977 *(± 0.001)*	0.983 *(± 0.001)*	0.441 *(± 0.007)*

TABLE 1. DL MODEL AVERAGE PERFORMANCE ON HQ AND LQ DATASETS. IN PARENTHESIS, THE STANDARD DEVIATION OF 5 INFERENCE TRIALS COMPUTED ON THE TEST SET.

Using the model's predicted scores, we generated a unified scale of 11 JOD, which represents a 5 JOD increase compared to the original HQ dataset. This enhancement not only represents the ability to evaluate more levels of quality but also to improve the precision of the resulting final scale.

Model validation: Results on simulated degradations.

Acutance on Dead Leaves vs Proposed model. The computation of the texture acutance on the Dead Leaves (DL) [3] chart has been widely used in the last years, and it is considered the standard procedure for measuring the camera texture quality rendition. In this section, we compare its performance to the proposed model. In all tests, the acutance is computed with fixed viewing conditions equal to the annotations on the realistic mannequins.

Increasing noise. This classic texture acutance is well known to be extremely sensitive to noise. In this experiment, the fine-grain was added to the luma channel (Y' in the Y'CbCR space). The variance of the noise depends linearly on the pixel intensity, and it is multiplied by a "noise gain" factor. The bigger the noise gain, the more visible the noise (see images in Figure 8 for an example).

As can be shown in Figure 8, given a high-quality image (DSLR 60Mpx), if we degrade it by adding fine-grain grey noise, the quality texture estimation drastically increases. During the acutance computation, we can mitigate the impact of the noise by estimating it on grey patches placed on the scene. However, since this method is generally circumvented by nowadays denoising algorithms that perform very well on flat regions, we decided to not use this mitigation.

In contrast, when taking an image of a Realistic mannequin on the same setup conditions as the DL chart, the ML-based model behaves differently. Unlike the acutance, a small quantity of fine noise tends to increase the texture perception. When the noise starts to look unnatural, the score decreases (see Figure 10). Thus, resulting in a significant correlation with human visual perception.

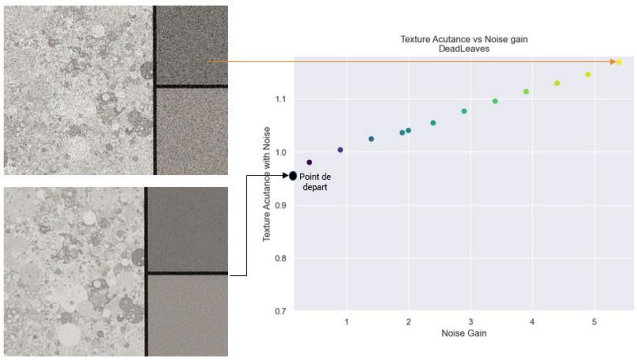

FIGURE 7. IMPACT OF ADDING NOISE TO THE SAME IMAGE, ON THE ACUTANCE MEASURE ON DEAD LEAVES CHART (SEE TEXT FOR MORE INFORMATION ON THE 'NOISE GAIN' PARAMETER)

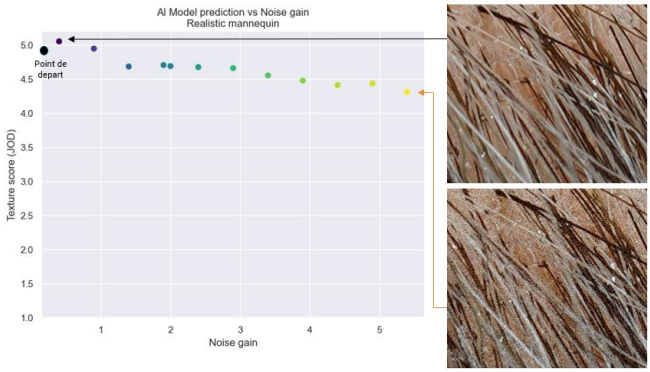

FIGURE 8. IMPACT OF ADDING NOISE TO THE SAME IMAGE, ON THE ML-BASED TEXTURE MEASURE ON REALISTIC MANNEQUINS

Similar results can be observed when simulating sharpening, using an unsharp masking (see Figures 10 vs 11) with the radius fixed to 10 pixels and varying strength (see "sharpening strength" in the Figures). As we increase the sharpening strength, the acutance on the DL augments, while the metric on the realistic mannequin decreases, as we expect by evaluating perceptually the images. We observe that the realistic mannequin measure is robust to sharpening while the acutance on the Dead Leaves is not.

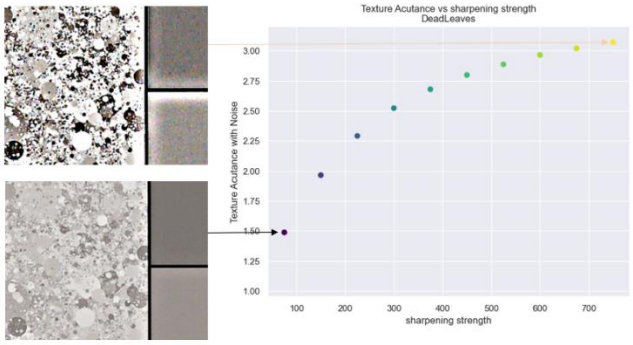

FIGURE 9. IMPACT OF ADDING SHARPENING TO THE SAME IMAGE, ON THE ACUTANCE MEASURE ON DEAD LEAVES CHART (SEE TEXT FOR MORE INFORMATION ON THE SHARPENING PARAMETER VALUES)

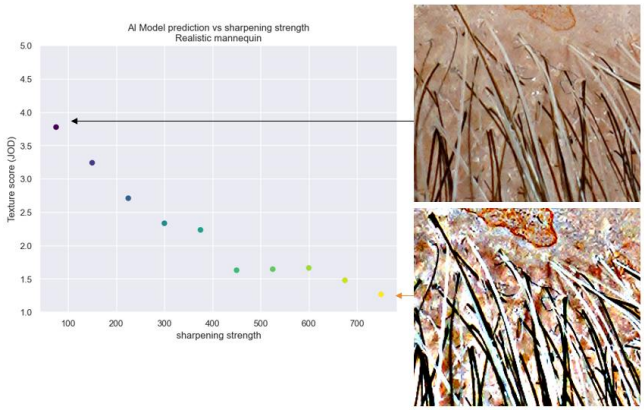

FIGURE 10. IMPACT OF ADDING SHARPENING TO THE SAME IMAGE, ON THE ML-BASED TEXTURE MEASURE ON REALISTIC MANNEQUINS (SEE TEXT FOR MORE INFORMATION ON THE SHARPENING PARAMETER VALUES)

Finally, we make the same comparison by applying resolution changes (see Figure 12 vs 13). We downscale, using Lanczos interpolation, the respective reference image to common smartphone resolutions, indicated by the reference dotted lines (1, 4, 6, 12, 24, 48, 60 Mpx). As expected, both metrics decrease with

the image resolution (i.e. increasing the downscaling factor). However, the Acutance on the DL chart saturates very fast, while the RM provides better quality resolution along the scale. The acutance computation is dependent on the viewing condition and distance, thus the plot in Figure 12 could be adapted to the LQ viewing conditions, but it implies having two different measures.

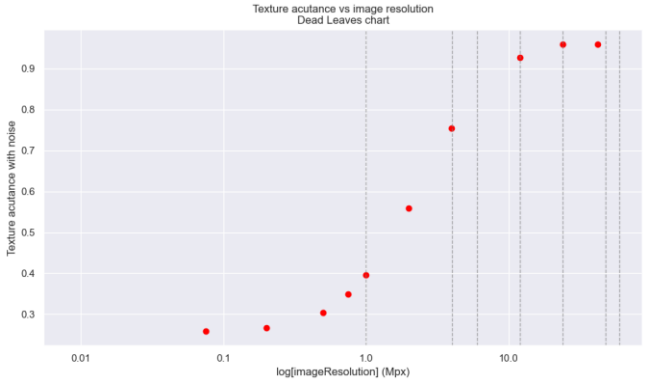

FIGURE 11. IMPACT OF CHANGING RESOLUTION TO THE SAME IMAGE, ON THE ACUTANCE MEASURE ON DEAD LEAVES CHART.

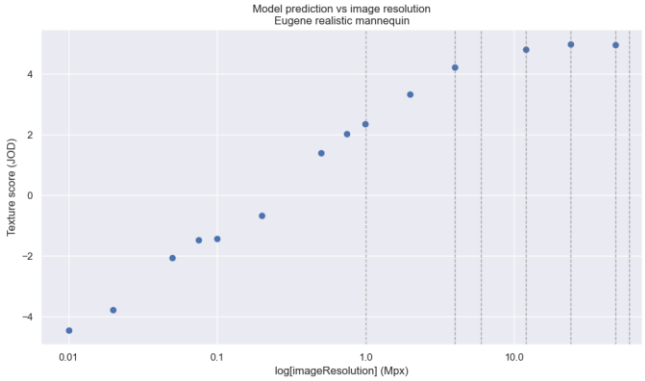

FIGURE 12. IMPACT OF CHANGING RESOLUTION TO THE SAME IMAGE, ON THE ML-BASED TEXTURE MEASURE ON REALISTIC MANNEQUINS.

This underscores the model's robust performance, sustained not only numerically through the high correlation between model predictions and ground truth but also perceptually in stress test scenarios. These scenarios, not included in the training, affirm that the model has learned the intrinsic nuances of natural texture rendering, in alignment with human perception.

Conclusions

This study introduces a new Texture quality estimation for Realistic mannequins with the following key attributes:

- Comprehensive scope: The method expands seamlessly from very low-quality inputs (e.g., video conference and doorbell images) to high-end captures from top-tier smartphones and DSLRs. The incorporation of a merging strategy, between high-quality and low-quality images, allows us to extend the evaluation range in a robust and precise way.

- Diverse model testing: Thorough evaluations were conducted on diverse models, including representatives of various demographics such as an old white male, a young Asian woman, and a deep-skinned woman. The results demonstrate consistently high performance across Regions of Interest (ROIs) when compared to perceptual annotations.

- Robustness: Notably, our proposed solution exhibits remarkable resilience to common challenges like noise and sharpening addition. This robustness addresses a classic problem encountered by other texture quality estimation methods, such as acutance on the Dead Leaves, strengthening the reliability and practicality of our approach in real-world applications.

References

[1] N Chahine, S Lahouar, S Soares, S Calarasanu, S Ferradans, B Pochon, F Guichard, "Noise quality estimation on portraits in realistic controlled scenarios", Society for Imaging Science & Technology, (2023).
[2] M Perez-Ortiz and R K Mantiuk. "A practical guide and software for analyzing pairwise comparison experiments" (2017).
[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] Gousseau, Yann, and François Roueff. *Modeling occlusion and scaling in natural images.* Multiscale Modeling & Simulation 6, no. 1 (2007): 105-134.
[6] van Zwanenberg, Oliver, Sophie Triantaphillidou, Robin Jenkin, and Alexandra Psarrou. "Edge detection techniques for quantifying spatial imaging system performance and image quality." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0-0. 2019
[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Conference on Computer Vision and Pattern Recognition, 2016. 2, 3, 6, 8, 9
[8] "IEEE standard for camera phone image quality," IEEE Std 1858-2016 (Incorporating IEEE Std 1858-2016/Cor 1-2017), 1–146 (2017).

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 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.*

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.*

Stefania Calarasanu *has earned a PhD in computer vision from the Pierre and Marie Curie University in collaboration with EPITA's research laboratory LRDE in 2015. She joined DXOMARK in 2019 and since then she works as an image quality engineer actively participating to the development of objective and perceptual quality metrics.*

Valentine Tosel *is a final year student at Ecole Polytechnique, France, on applied math and computer science. Her interests include probability, statistics and its application to imaging problems.*

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). Since December 2020, he is working full time at DXOMARK Image Labs as a Ph.D. student in collaboration with INRIA Paris. His work focuses on automated image quality assessment.

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.*