https://doi.org/10.2352/EI.2023.35.8.IQSP-316

Noise quality estimation on portraits in realistic controlled scenarios

Nicolas Chahine, Sofiene Lahouar, Samuel Soares Santos, Stefania Calarasanu, Sira Ferradans, Benoit Pochon, Frédéric Guichard.
DXOMARK, Boulogne-Billancourt, France.

Abstract

The wide use of cameras by the public has raised the interest of image quality evaluation and ranking. Current cameras embed complex processing pipelines that adapt strongly to the scene content by implementing, for instance, advanced noise reduction or local adjustment on faces. However, current methods of Image Quality assessment are based on static geometric charts which are not representative of the common camera usage that targets mostly portraits. Moreover, on non-synthetic content most relevant features such as detail preservation or noisiness are often un-tractable.

To overcome this situation, we propose to mix classical measurements and Machine learning based methods: we reproduce realistic content triggering this complex processing pipelines in controlled conditions in the lab which allows for rigorous quality assessment. Then, ML based methods can reproduce perceptual quality previously annotated. In this paper, we focus on noise quality evaluation and test on two different setups: closeup and distant portraits. These setups provide scene capture conditions flexibility, but most of all, they allow the evaluation of all quality camera ranges from high quality DSLRs to video conference devices. Our numerical results show the relevance of our solution compared to geometric charts and the importance of adapting to realistic content.

Introduction

In the past few years, the use of videoconferencing has exploded. This builds on the widespread usage of selfie photos so common in social media, giving a lot of attention to portrait rendition. To match the expectations of the market, camera providers are developing dedicated portrait pipelines that apply a different digital treatment to the face region and to the background. Therefore, camera evaluation techniques cannot continue to be the same: we need new evaluation methods that take into account the context of the scene.

Deadleaves [9][10] and OECF charts have proven to be rich charts to evaluate texture and noise attributes for classical cameras where the image processing pipeline is well defined and understood, such as DSLR cameras in manual mode for instance. With such charts and cameras, analytical noise estimation methods such as the ISO15739 Visual Noise can be computed on uniform patches and correlate well with perceptual quality assessment of noise on various content, under the same capture conditions. However, we know that current smartphone cameras are using more complex pipelines that adapt to the content of the scene, that is to say, the capture parameters and digital algorithms applied will not be the same for a portrait, a natural scene or synthetic geometrical content, even under the same light conditions.

Uniform Patches vs Portraits. In Figure 1, we show how the same devices behave differently regarding noise rendition depending on the content of the scene. For every device, we capture with the selfie camera, under the same light conditions (100lux and 20Lux, 45 cm distance from the object) the Deadleaves chart and two realistic mannequins. We crop the relevant areas on each image, that is, flat regions for the Deadleaves and the forehead for the mannequin (see Figure 2). To compare fairly, we select for each image the uniform crop on the Deadleaves chart with lightness L* the closest possible to the lightness L* of the crop of the forehead of the mannequin. On the Deadleaves crop we compute the Visual Noise mapped to a JND of noisiness scale such as described in [8]. The lower the value the lower the perceived noise. On the other hand, we annotate the image quality noise rendition on the mannequin's crop in quality units ('Just Objectionable differences', see next section for more details). Viewing conditions during the annotations are the same (image viewed with a cutoff frequency of 30 cycles per degree). In the noise annotation on the realistic mannequins, the lower the value the greater the perceived noise. The perfect correlation line between these two metrics is displayed in grey. The blue and orange dots show the camera quality range of different devices at respectively 100Lux and 20Lux. Note that since the viewing conditions and the scale of annotations are similar, the slope of the grey line is one (absolute value), but the offset is arbitrary.
As we can see, most devices are aligning on the line, meaning that the relative perception of noise on the forehead of the mannequin and on a uniform patch is similar. However, some devices are a bit further from the line, which indicates that the devices have different noise rendition depending on the content of the scene.

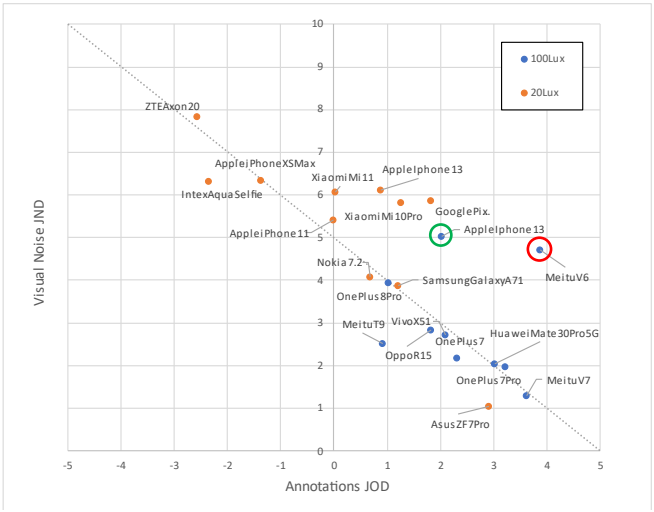

Figure 1. Noise perceptual evaluation on portraits vs noise estimation on the Deadleaves flat patches under the same viewing

conditions. The grey line shows the perfect correlation between both measures.

Let's observe the different case for these non-aligned devices. In Figure 2, we can take the extreme case of MeituV6 (in a red circle on Figure 1). On the flat patches of the Deadleaves, we can observe noise, but this amount of noise is not visible on the mannequin which has been extremely denoised.

On the other hand, in Figure 3, the images produced by the Device A (yellow circle in Figure 1), are fairly clean on the Deadleaves flat patches (see upper image in Figure 3) but present over sharpened noise artifacts that were considered as more objectionable on the forehead crop (see lower image).

Finally, in Figure 4, we can observe the images produced by the iPhone13 (green circle) on the two charts: noise is quite visible and objectionable on the flat patches of the Deadleaves whereas it was judged much more acceptable on the forehead, probably because of its nature: a fine grain luminance noise which does not interfere with mental representation of fine details of the skin on the mannequin.

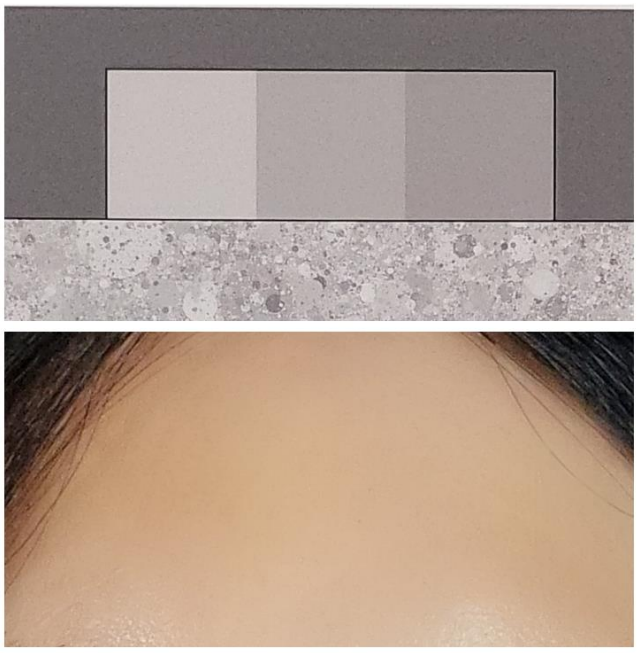

Figure 2. MeituV6 capture of the Deadleaves chart (upper image) and a crop of Sienna realistic mannequin (lower image). Both crops are 800 by 400 pixels. Even though the images were captured in the same conditions, the Deadleaves chart displays much more noise. This illustrates how the noise rendition depends on the content of the scene.

This motivates us to develop content-aware noise quality estimations based on perceptual evaluation of such type of images. In order to adapt to this diverse content, we propose to tackle the perceptual evaluation of noise using Machine Learning (ML).

In this work, we focus on realistic mannequin setups (see figures 2 and 3) that are representative of a real-life portrait situation. The advantage of using mannequins instead of real human models is that the capture conditions can be fully reproduced.

As for all ML solutions, we need to construct a perceptually based annotated data set of images, which is highly challenging since it needs to cover all the camera quality ranges. Moreover, evaluating noise on textured areas (namely face attributes) is not evident since its perception can change depending on the size of the image and the viewing conditions.

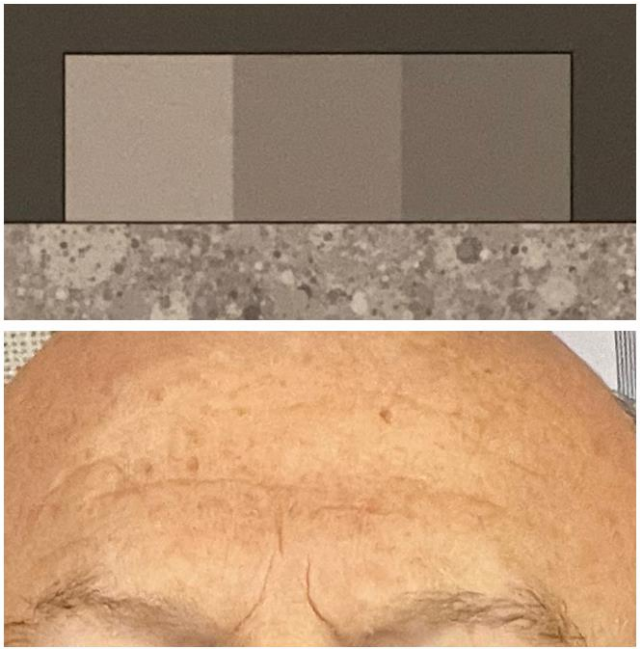

Figure 3 Apple iPhone13 images taken under the same conditions. Noise is visible on the flat patches of the Deadleaves (upper image) and the forehead of the mannequins (lower image). On the forehead, the noise has been judged as less objectionable, probably due to its fine grain luminance nature that does not interfere with the expected texture of the skin.

Novelty. The use of ML models for noise quality assessment was proposed in [5] although they did not study the case for portraits and 3D mannequins which is an important use case due to the increase of video conferencing. The use of ML models to estimate the image quality was also used for realistic mannequins in the context of detail preservation in [6].

This paper proposes a new measure for noise quality assessment in realistic controlled portrait scenarios based on Machine Learning methods. This problem is of fundamental importance given the current popularity and impact of selfie pictures and videos in social media, as well as the increased usage of video conferencing.

Method

The first step in the construction of an ML-based solution is always the creation of a relevant database for training and testing. The training data set allows the algorithm to learn from the examples, while the testing set provides an estimate of how well the algorithm behaves in production, that is, on unseen images during training. The content of the database implicitly defines the requirements of the model.

Image database construction. The goal of this study is the assessment of perceptual noise on realistic mannequins' setups on

an extended range of image qualities. This range of qualities is characterized by the introduction of different quality devices such as DSLRs, smartphone and video conference (i.e. laptop camera, tablets) in the image dataset, but also the use of variable shooting conditions (i.e. distance, lighting condition). Since we want to observe noise on a scale dependent content (a face), and compare devices between each other's with equivalent framing, the annotation of images requires a re-scaling so that the contents are displayed at the same size. However, applying strong rescaling factors to an image with low resolution is sometimes not meaningful, as it will also impact the visual perception of noise, leading for example to trivial image comparisons.

This makes the perceptual evaluation of noise a complex and challenging task. To overcome trivial comparisons between images of different frequency levels, we propose to split our database into 2 subsets with respect to the size of the face in the image. This categorization can also be interpreted as a partition between two viewing conditions, one for high quality (HQ) images and the other for low quality (LQ). We propose to assess the noise on different regions of interest: for high quality images, we will focus on the forehead area of the realistic mannequins, while on the lower quality subset, we will rather concentrate on the whole face area in order to maximize the frequency information pertinent to evaluating the perceived noise (see Figure 4 and 5 for an example).

Figure 5. Example of Low-Quality Video Conference samples. From the complete setup (upper image) two Regions of Interest (ROI) are extracted, one for each mannequin.

Figure 6. Example of a pairwise comparison task done by annotators.

Perceptual annotations and scale quality construction. To obtain quality labels, we need to collect perceptual evaluations. The goal is to generate a psychophysical image quality scale. The observer is asked to estimate the image quality by analyzing its amount of noise, compared to a mental representation of the ideal image. However, evaluating the image quality in an absolute manner (without reference or comparisons) is a complicated task, and usually requires the opinion of a large group of observers. One of the most precise approaches is pairwise comparisons, where one can infer quality scores using a comparison matrix. In this context, an annotator is presented with two images, and he/she needs to choose which one has a better quality regarding noise (see Figure 6 for an example). The main problem with this approach is its quadratic growth, which means that the cost and difficulty increases in a quadratic manner with the size of the dataset. Following standard recommendations ITUT [1] and ITU-R [2], a minimum of 15 full comparisons O(n2) (i.e., 15 annotators to compare all $n(n-1)/2$ pairs) is required to generate reliable results, which can be costly and time consuming. Usually, it is not possible to achieve a full design (a full passage on all data points for each observer), which implies the need for a sampling strategy or active sampling as it's referred to in the literature. We adapt an active sampling technique based on information gain [3], in order to extract the most interesting pairs on each iteration and make the comparison task the most efficient possible. The score inference is finally done using TrueSkill algorithm, from which we generate scores on a JOD (Just Objectionable Difference) scale, based on Thurstone case V observer model [4]. A JOD is a quality unit obtained statistically from annotations. Image *A* is one JOD apart from Image *B* if 75% of the annotators agree that the quality of image A is greater than image B [7].

Technical details of the annotated database. In Table 1, we show the final content of the four constructed databases: high quality and low quality for two different mannequins (Eugene and Sienna, see Figure 5 and 6 for more details). The train/test split was defined in such a way that there are no overlapping devices, the reason being that we would like to understand the capacity of the model to extend to unseen cameras once the model is put in production. Moreover, the devices in train and test were chosen to maximize the rendition diversity. As an example, in Figure 7, we observe the empirical distribution of the database's image quality scale for Eugene HQ and Eugene LQ. The scale definition is invariant to a constant offset. Without loss of generality, we have

set for each scale, one sample to 0 (a sample is an image of one device at a fixed lighting condition).

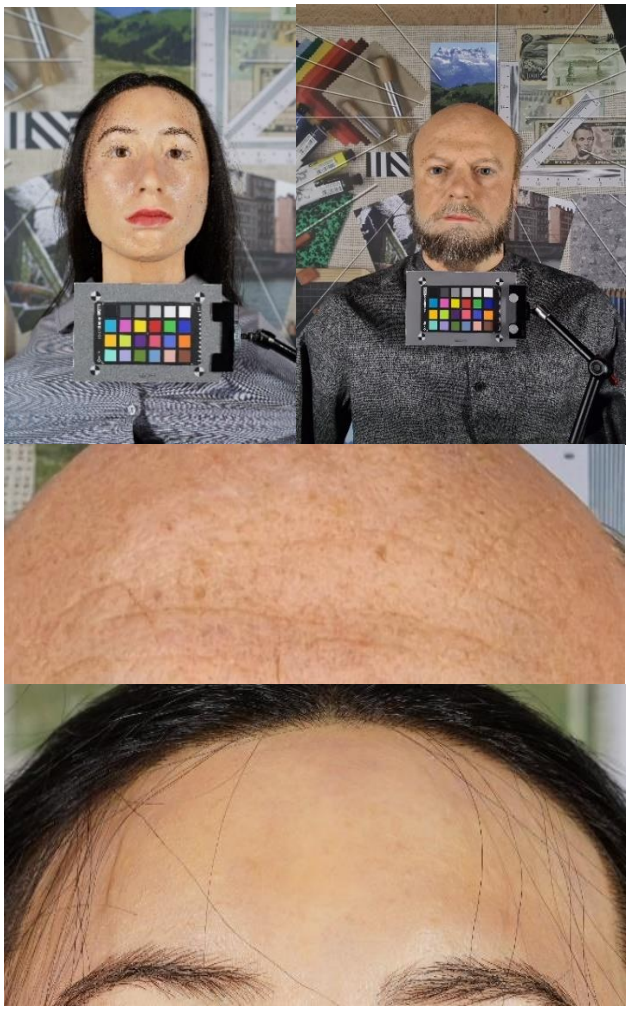

Figure 6. Example of Average/High-Quality samples, source images (upper image) and extracted ROIs (lower images).

Composition of the data set		
	#Images Train/Test	**#Different Devices** Train/Test
HQ Eugene	253/48	60/23
HQ Sienna	190/48	59/23
LQ Eugene	232/58	51/10
LQ Sienna	178/39	18/10

Table 1. Content of the constructed database, see text for more details.

ML models and validation. The previous annotated datasets allow us to train two CNN-based regressors that estimate the noise characterization of the image, one per quality level. Experimental results show that we attain better performance with a multitask setting that predicts not only the JOD quality level of the input image, but also classifies its content (Sienna vs Eugene). Our hypothesis is that, with a common model, during training, the deep learning model has access to more examples: those of Eugene and Sienna, instead of only one mannequin.

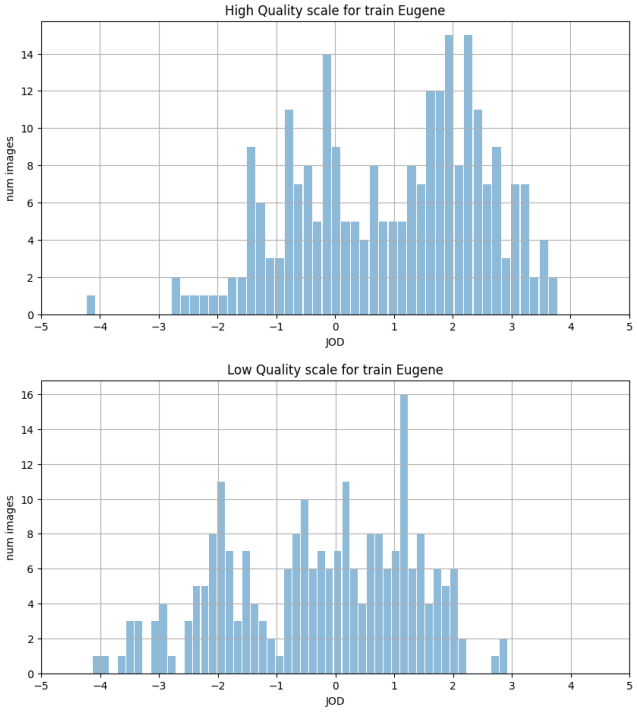

Figure 7. Histograms of the training set for the Eugene databases, see text for more details.

Results

The two proposed models were applied to the testing set. In Table 2, we present the Spearsman rank correlation coefficient (SROCC) for the 4 testing sets. The SROCC indicates if the ground truth and the model output rank similarly the input images (the closer to 1 the better).

SROCC *Results on testing set*			
Model Eugene HQ	**Model Sienna HQ**	**Model Eugene LQ**	**Model Sienna LQ**
0.95	0.88	0.88	0.71

Table 2. Spearsman's rank correlation coefficient for the two models (HQ and LQ) evaluated on the corresponding testing set.

As we can see this is actually the case. This can also be observed in Figure 8. The results are better for the Eugene mannequin, probably due to the bigger size of the training set.

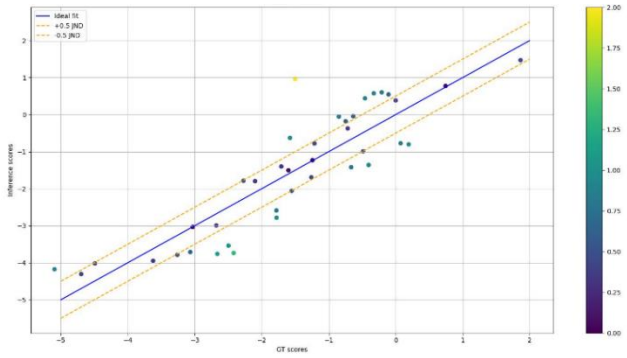

Figure 8. Annotated ground truth JOD scores vs predicted JOD scores for the LQ model on the Eugene and Sienna testing set. The blue line indicates the perfect prediction, and the two yellow lines indicate a 0.5 JOD deviation. The point colors indicate the error, ie the distance to the ground truth. As we observe, most of the samples are within the (-0.5,+0.5) interval.

We also tested the ML setup with four different models (one per mannequin and quality level) and without multitasking (*i.e.* a single regression output that corresponds to the estimated quality of the input image), we can see the results on Table 3. Comparing Table 2 and Table 3 we observe that using multi-tasking improves mostly the Sienna LQ results.

SROCC *Results on testing set*		
	Model HQ	**Model LQ**
Eugene	0.95	0.88
Sienna	0.92	0.88

Table 3. Spearsman's rank correlation coefficient for the four models (one per mannequin and quality level) evaluated on the corresponding testing set.

Interpretation. This suggests that, in the multitask scenario, the features generated by the CNN are mutualized for the LQ model between Sienna and Eugene, whereas for the HQ model, this is not the case. Indeed, locally, the foreheads of both mannequins contain different fine details that are only visible in the high-quality viewing conditions. In the low-quality viewing conditions, where we look at the whole face, and where fine details are less visible, the model can leverage the common face features.

Conclusions

In this paper we proposed an experimental setup that allows to automatize the perceptual analysis of noise rendition on realistic portraits. We have constructed a database using pairwise comparisons and the TrueSkill algorithm to generate a JOD quality level for each image. This ground truth was then used to train two CNN-based regressor models. The numerical results confirm the validity of our approach for cameras never seen during training, showing that using Machine Learning to estimate the perceptual image quality of mannequin setups is a robust choice to tackle the automatic noise evaluation.

References

[1] ITU-T RECOMMENDATION, P. "Subjective video quality assessment methods for multimedia applications." International telecommunication union (1999).

[2] BT, RECOMMENDATION ITU-R. "Methodology for the subjective assessment of the quality of television pictures." International Telecommunication Union (2002).

[3] Active Sampling for Pairwise Comparisons via Approximate Message Passing and Information Gain Maximization

[4] Murphy, Kevin P. Machine learning: a probabilistic perspective. MIT press, 2012.Author Biography

[5] S. Belkarfa, AH Choukarah, M Tworski. "Automatic Noise Analysis on still life chart". In London Imaging Meeting (Vol. 2021, No. 1, pp. 101-105). Society for Imaging Science and Technology.

[6] 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

[7] M Perez-Ortiz, A Mikhailiuk, E Zerman, V Hulusic, G Valenzise, and R K Mantiuk. *From pairwise comparisons and rating to a unified quality scale*. IEEE Transactions on Image Processing, 29:1139–1151, 2019.

[8] Thomas Bourbon, Coraline Hillairet, Benoit Pochon, Frederic Guichard. *New visual noise measurement on a versatile laboratory setup in HDR conditions for smartphone camera testing*. In Electronic Imaging 2022, EI.2022.34.9.IQSP-313.

[9] Gousseau, Yann, and François Roueff. *Modeling occlusion and scaling in natural images.* Multiscale Modeling & Simulation 6, no. 1 (2007): 105-134.

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

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.*

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.*

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*

Sofiene Lahouar *is a machine learning and data science graduate. He completed a double degree program between the Higher School of Communication of Tunis (SUP'COM) and Telecom Paris, graduating in 2022.*

Samuel Soares Santos *is a deep learning and computer vision engineer. He obtained his undergraduate degree from the University of Brasília and followed a Master's double degree program between the University of Brasília and the Polytechnic Institute of Bordeaux (2019-2021).*
Since 2022, he has been working full time at Parrot Drones.