



OPEN ACCESS

EDITED BY

Yong Hu,
The University of Hong Kong, Hong Kong SAR,
China

REVIEWED BY

Liya Ding,
Southeast University, China
Jing Teng,
North China Electric Power University, China

*CORRESPONDENCE

Camille Simon-Chane
✉ camille.simon-chane@ensea.fr

RECEIVED 10 May 2023

ACCEPTED 24 July 2023

PUBLISHED 15 August 2023

CITATION

Brémond-Martin C, Simon-Chane C,
Clouchoux C and Histace A (2023) Brain
organoid data synthesis and evaluation.
Front. Neurosci. 17:1220172.
doi: 10.3389/fnins.2023.1220172

COPYRIGHT

© 2023 Brémond-Martin, Simon-Chane,
Clouchoux and Histace. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).
The use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in this
journal is cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Brain organoid data synthesis and evaluation

Clara Brémond-Martin^{1,2}, Camille Simon-Chane^{1*},
Cédric Clouchoux² and Aymeric Histace¹

¹ETIS Laboratory UMR 8051 (CY Cergy Paris Université, ENSEA, CNRS), Cergy, France, ²Witsee, Neoxia, Paris, France

Introduction: Datasets containing only few images are common in the biomedical field. This poses a global challenge for the development of robust deep-learning analysis tools, which require a large number of images. Generative Adversarial Networks (GANs) are an increasingly used solution to expand small datasets, specifically in the biomedical domain. However, the validation of synthetic images by metrics is still controversial and psychovisual evaluations are time consuming.

Methods: We augment a small brain organoid bright-field database of 40 images using several GAN optimizations. We compare these synthetic images to the original dataset using similitude metrics and we perform a psychovisual evaluation of the 240 images generated. Eight biological experts labeled the full dataset (280 images) as synthetic or natural using a custom-built software. We calculate the error rate per loss optimization as well as the hesitation time. We then compare these results to those provided by the similarity metrics. We test the psychovalidated images in a training step of a segmentation task.

Results and discussion: Generated images are considered as natural as the original dataset, with no increase of the hesitation time by experts. Experts are particularly misled by perceptual and Wasserstein loss optimization. These optimizations render the most qualitative and similar images according to metrics to the original dataset. We do not observe a strong correlation but links between some metrics and psychovisual decision according to the kind of generation. Particular Blur metric combinations could maybe replace the psychovisual evaluation. Segmentation task which use the most psychovalidated images are the most accurate.

KEYWORDS

psychovisual, metric, validation, brain organoid, AAE

1. Introduction

The scarcity of public datasets of annotated biomedical images remains an unresolved bottleneck to develop specialized and robust analysis tools. Often, research groups do not share experimental data for privacy reasons. The high costs of equipment, long acquisition times, and necessary in-depth expertise can be a brake to acquisitions by other teams (Chakradhar, 2016). To benefit from the advances in deep-learning (DL) for automated image analysis, large training datasets are necessary. Moreover, original dataset constraints create a problem of class imbalance with deep learning training procedures. These problems are emphasized with small sets, reduced to a few images (Tajbakhsh et al., 2016).

Data augmentation is widely used in the biomedical domain to increase the size of image datasets (Singh and Raza, 2021). While classical data augmentation, based on transformations (flip-flops, rotation, whitening, etc.), does not increase the diversity of the dataset, a solution widely provided in the biomedical field is the use of Generative Adversarial Networks (GAN) which produce new synthetic images from natural ones (Yi et al., 2019; Lan et al., 2020).

GANs are unsupervised deep-based architectures composed of generator and a discriminator. The generator aims at creating visually realistic and natural images while the discriminator tries to decipher whether the result is generated (Goodfellow et al., 2014). Both networks are trained simultaneously with the same loss function. Since its first creation, multiple GAN architecture variations have been proposed to generate and extend biomedical datasets (Lan et al., 2020; Fernandez et al., 2021; Chen et al., 2022). However, the validation of these synthetic images remains a challenge Alqahtani et al. (2019). Two main evaluating methods are existing the non-automated based upon psychophysics methods in perceptual psychology, and automated methods based upon metrics (Salimans et al., 2016a; Zhou et al., 2019). Psychovisual evaluation is a time consuming gold standard which requires many subjects to reduce its' intrinsic subjectivity. On the other hand, there is no commonly approved specific metric to evaluate whether GAN-generated synthetic images can be considered as natural. The use of common metrics is also controversial (Borji, 2019).

We consider the case of brain organoids, three dimensional cultures differentiated from pluripotent stem cells (Lancaster et al., 2013). After a neural induction, brain organoids present cell communications, organized tissues and an organization similar to the brain with various regions such as neuro-epithelial zones (Kelava and Lancaster, 2016). They complement *in vivo* brain models to follow physiological and pathological brain development. However, brain organoids suffer from batch syndrome: in the same culture environment they do not innately develop comparable morphologies (Lancaster et al., 2013).

There are no specific tools to study the development of brain organoids. Though it may seem natural to implement machine learning algorithms to aid this task, few brain organoid image datasets are publicly available. Between January 2018 and June 2020 for instance only six over 457 articles in the brain organoid field let the image datasets in openaccess and only one concerns brain organoids in bright-field (Brémond Martin et al., 2021). The emergence of brain organoids has created a new field, few laboratories have the knowledge and experience necessary to grow these cultures. Pandemic restrictions have further limited the possibility for several teams to grow and image such culture over the past few years.

The largest public brain-organoid image dataset we know of contains 40 images (Gomez-Giro et al., 2019). Data augmentation solutions have already been used to increase the size and diversity of this brain organoid bright-field dataset. An adversarial autoencoder (AAE) seems the architecture the most suited to augment images of brain organoid bright-field acquisition (Brémond Martin et al., 2022a). This AAE differs from the original GAN architecture by the input given to the encoding part (original images) and its generative network containing an auto-encoder-decoder framework (Goodfellow et al., 2014; Makhzani et al., 2016). The encoder learns to convert the data distribution to the prior distribution, while the decoder learns a deep generative model that maps the imposed prior to the data distribution thanks to a latent space (Makhzani et al., 2016). As is typical with AAEs, the images generated are visibly blurry.

To improve the sharpness during the generation, we test various loss functions to improve the adversarial network

(Brémond Martin et al., 2022a). However, these results are based upon metric calculation and a dimensional reduction to compare all feature images (original and generated with each optimization) in the same statistical space. Indeed some metrics may not be suited to identify the naturalness of an image as they are originally created to test the similitude or the quality of images (Borji, 2019; Brémond Martin et al., 2022a). In our previous contribution we observe the data augmentation strategy based upon AAE loss optimizations used during the training step of a DL segmentation algorithm improve the quality of the shape extraction of brain organoids (Brémond Martin et al., 2022a). The first raised question is does these images seem as natural as the original images to furnish a better segmentation quality compared to a result issue from classic data augmentation strategies? Another fundamental issue remains unresolved: do these synthetic images seem natural to a biological expert point of view as for metric(s)? Psychovisual evaluations have been already made on others bright-field cell synthetic cell generation (Malm et al., 2015). This evaluation is an important step for the validation of a particular generative model of images. Thus the selected images as natural by Human Biological experts could maybe help to train deep based segmentation methods and characterize their development but with now a double psychovisual-metric validation.

We propose to evaluate the synthetic images generated by an AAE (Brémond Martin et al., 2022a) using both with similarity metrics and biological experts. The purpose of this article is to give the lacking non-automated psychovisual evaluation of the synthetic images which is a new contribution compared to our previous contribution which focus on automated metric based and statistical strategies in order to understand if the naturalness of these images could explain the segmentation results (Brémond Martin et al., 2022a). The second original part of the work is to find a metric combination which may replace or complete the psychovisual non-automated evaluation. Related work is presented in the following section. Section 3 describes the generative network, the metric evaluation and the psychovisual evaluation. Section 4 successively shows the results for the metric evaluation and the psychovisual evaluation, followed by a cross-analysis of the two evaluation methods. These results are analyzed in the Section 5. Section 6 sums up the main contributions of this paper.

2. Related work

The first generative adversarial network, proposed by Goodfellow et al. (2014), is constituted by two connected networks: the generative model (*GM*) maps the images into the space (z) by an objective function (F); the discriminative model (*DM*) determines the probability for which a point from z belongs to the original dataset (o) or to the generated dataset (g). Training F increases the probability that the data synthesized is attributed to o . The probability of correct sample labeling (belonging to generated g or original o) is maximized by D . Simultaneously, GM is trained to leverage the discriminator function expressed by:

$$\min_{GM} \max_{DM} F(DM, GM) = E_o P_{data}[\log D_o] + E_g P_g[\log(1 - D(G_z))]. \quad (1)$$

An overview of various GAN architectures used for medical imaging is given by Yi et al. (2019). Previous work compares five GAN architectures commonly used for biological datasets to find the best suited network to increase a small database of bright-field images of organoids (Brémond Martin et al., 2022a): the original GAN implementation (Goodfellow et al., 2014); CGAN gives to the generator input the correct label (physiological or pathological) (Mahmood et al., 2018); DCGAN by Radford et al. (2015) is constituted by a convolutional neural networks instead of the generator; INFOGAN uses the generated images at an epoch to train the subsequent (Hu et al., 2019); AAE by Makhzani et al. (2016) uses an auto-encoder as a generator. This work evaluates the generated images using metrics.

As already reported in the literature by Lan et al. (2020), GAN and CGAN produce mode collapse with such a small dataset. In addition, GAN produces a white imprint around the shape of the organoid. DCGAN and INFOGAN generate a divergent background which makes the images difficult to exploit. This is probably due to the high variability of the small dataset, as mentioned in Lan et al. (2020). Only AAE produces images of similar quality to the original and seems the best architecture to generate images with a few input. However, AAE generates blurry images and the background differs from the light-to-black gradient present in many bright-field images.

Noise-injection during the AAE generation was studied to improve the background generation with satisfying results in Brémond Martin et al. (2022b). In this paper, we further explore the effect of loss optimization on the AAE architecture to improve the sharpness of the generated images. The results are evaluated by an automatic approach based on metrics and by a psychovisual study. The last part aims at giving an application of such psychovalidated images in a segmentation task.

3. Methods

In this section, we first present the generative methods and optimizations used. We follow by a description of the metric and psychovisual evaluations including the experimental setup, datasets and experts. We present the analysis methodologies of the neurobiologist decisions and the comparison of psychovisual evaluations with metric calculations. We finish by giving an asset of the importance of psychovalidated images in an augmented dataset strategy for a segmentation task.

3.1. Generative adversarial networks

3.1.1. Original images

Our dataset is composed of 40 microscope acquisitions and 240 synthetic images created by a AAE loss optimizations. “Original” images are the 40 bright-field brain organoid acquisitions provided by Gomez-Giro et al. (2019). This dataset is made of image of 20 physiological and 20 pathological organoids acquired over three days on the same apparatus. These input images ($1,088 \times 1,388$ pixels) are cropped and scaled to 250×250 pixels, maintaining their original proportions. A scale factor of 4 is chosen so the scripts can

run in a reasonable amount of time without downgrading the input image quality too much.

3.1.2. Loss optimizations

The images generated by the AAE architecture we use are somewhat blurry. To overcome this phenomenon we study how the discriminator loss can influence the quality of the image generation. We consider six losses: the Binary Cross Entropy (BCE) which is most commonly used in GANs and five other losses which are specifically known to improve the contrast or sharpness of the generated images.

BCE is the most commonly used loss for GANs and the baseline of this work. It is calculated by:

$$\text{BCE} = -\frac{1}{n} \sum_{i=1}^n (y_i(\log(y'_i)) - ((1 - y_i)(\log(1 - y'_i))) \quad (2)$$

with y the real image tensor and y' the predicted ones (Makhzani et al., 2016) and n the number of training.

Summing the L1 norm to the BCE is reported to reduce overfitting (Wagnier-Dauchelle et al., 2019). We hypothesize this norm could improve the quality of the generation as reported in image restoration tasks which does not over-penalize large errors (Zhao et al., 2017).

$$\text{L1} = \frac{1}{n} \sum_{i=1}^n |y_i - y'_i|_1 \quad (3)$$

$$\text{BCE}_{\text{L1}} = \text{BCE} + \alpha \text{L1}. \quad (4)$$

α is set to 10^{-4} , as in the original paper.

The least square loss (LS) is reported by Mao et al. (2017) to avoid gradient vanishing in the learning process step resulting in better quality images:

$$\text{LS} = \frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2 \quad (5)$$

A Poisson loss is used in Wagnier-Dauchelle et al. (2019) to improve the sensitivity of a segmentation task:

$$L_{\text{Poisson}} = \frac{1}{n} \sum_{i=1}^n (y'_i - y_i) \log(y'_i + \epsilon) \quad (6)$$

where ϵ is a regularization term set to 0.25.

The DeblurGAN was developed to unblur images using the Wasserstein loss (Kupyn et al., 2018). Since we are also interested in deblurring the output images, we have tested this loss with the proposed AAE where $(P(y, y'))$ is the joint distributions of y and y' for which the distributions are equal to P_y and $P_{y'}$, and $p(y, y')$ the proportion of y or y' to move to have $P_y = P_{y'}$:

$$W_{\text{ass}}(P(y, y')) = \sum_{i=1}^n \inf_{p \in P(y, y')} E_{y_i, y'_i} \delta p(\|y_i - y'_i\|) \quad (7)$$

However, we do not apply a l2 content loss such as in Kupyn et al. (2018) added to the Wasserstein loss, or add a penalty gradient

to the Wasserstein loss such as in [Gulrajani et al. \(2017\)](#). Our dataset containing various subclasses (physiological and pathological brain organoid images acquired at three developmental stages), we aim at creating a Normalized Wasserstein loss to avoid imbalanced mixture proportions ([Balaji et al., 2019](#)). We apply the L2 normalization on the Wasserstein loss, producing a new loss we call the Perceptual Wasserstein loss for the first time to our knowledge, applied on an AAE architecture:

$$P.Wass(P(y, y')) = \sum_{i=1}^n \sqrt{\inf_{p(y_i, y'_i)} E_{y_i, y'_i} \delta p(\|y_i - y'_i\|)^2} \quad (8)$$

3.1.3. Training

[Figure 1](#) shows the global training setup. The 40 original images are used to generate 40 synthetic images for each architecture (each loss). Input and output images measure 250 × 250 pixels. Training lasts 2,000 epochs for each optimization; this corresponds to the plateau before over-fitting for all loss optimizations.

3.1.4. Resources

The GAN algorithms are developed in Python 3.6 with an Anaconda framework containing the 2.3.1 Keras and 2.1 Tensorflow versions. All scripts are executed on an Intel Core i7-9850H CPU with 2.60 GHz and a NVIDIA Quadro RTX 3000s GPU device.

3.2. Metric evaluation

Six metrics are used to compare the similitude of the synthetic images generated by the AAE to the original dataset. A blur metric is used to evaluate the quality of these synthetic images.

The Fréchet Inception Distance (FID) is calculated between two groups of images ([Heusel et al., 2017](#)). This score tends toward low values when the two groups (original O or generated G images) are similar, with μ the average value of the pixels of all images of a group, and Σ the covariance matrix of a group:

$$FID(O, G) = \|\mu_O - \mu_G\|^2 + \text{Tr}(\Sigma_O + \Sigma_G - 2(\Sigma_O \Sigma_G)^{\frac{1}{2}}) \quad (9)$$

The Structural Similarity Index (SSIM) is calculated using luminance, contrast and structure by [Wang et al. \(2004\)](#) between two images o and g belonging respectively to O and G .

$$SSIM(o, g) = \frac{(2\mu_o\mu_g + c_1)(2\sigma_{og} + c_2)}{(\mu_o^2 + \mu_g^2 + c_1)(\sigma_o^2 + \sigma_g^2 + c_2)} \quad (10)$$

where σ represents the standard deviation, c_1 is a constant that ensures the luminance ratio is always positive when the denominator is equal to 0, and c_2 is an other constant for the contrast stability. The SSIM ranges between 0 (no similitude) and 1 (high similitude).

The Universal Quality Metric (UQM) is based on the calculation of the same parameters as SSIM and was proposed by

[Wang and Bovik \(2002\)](#). UQM ranges between 0 and 1 (1 being the highest quality):

$$UQM(o, g) = \frac{4\mu_o\mu_g\mu_{og}}{(\mu_o^2 + \mu_g^2)(\sigma_o^2 + \sigma_g^2)} \quad (11)$$

Entropy-based Mutual Information (MI) measures the correlation between original and generated images and ranges between 0 (no correlation) and 1 (high correlation) ([Pluim et al., 2003](#)):

$$MI(o, g) = \sum_{o \in O} \sum_{g \in G} P(o, g) \log \frac{P(o, g)}{P(o)P(g)} \quad (12)$$

where $P(o, g)$ is the joint distribution of o belonging to O and g from G .

The Mean Square Error (MSE) between an original image and a synthetic image is calculated as:

$$MSE(o, g) = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n (o(i, j) - g(i, j))^2 \quad (13)$$

The Peak Signal to Noise Ratio (PSNR) indicates a high signal power against noise, as used in [Jiang et al. \(2021\)](#). High values correspond to qualitative images. Pixels in images are ranked between 0 and 255, thus the maximum pixel value of an image is noted $\max(o)$ and equals at most 255.

$$PSNR(o, g) = 20 \log \max(o) - 20 \log MSE(o, g) \quad (14)$$

where \log denotes the common logarithm.

As a quality metric we calculate the blur index proposed by [Tsomko et al. \(2008\)](#) based on local image variance. A low score corresponds to a sharp image. In the following equation, the image is of size (m, n) , the predictive residues a given image pixel are $p(i, j)$ and their median $p'(i, j)$:

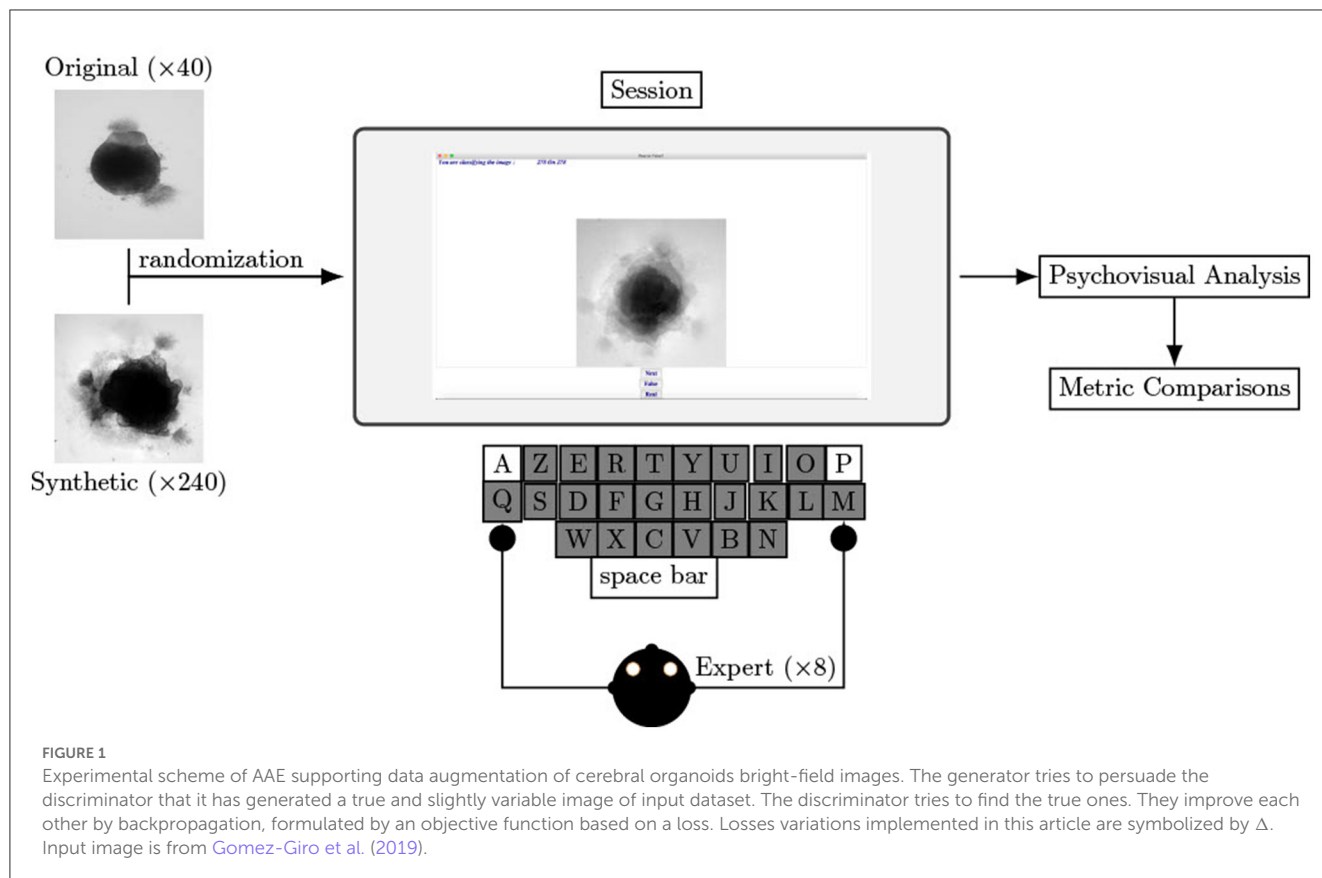
$$\text{Blur} = \frac{1}{m(n-1)} \sum_{i=1}^m \sum_{j=1}^{n-1} [p(i, j) - p'(i, j)]^2 \quad (15)$$

The FID is designed to compare groups of images. We thus successively compare each group of synthetic images with the original input images. The FID reference range is calculated on the original image developmental stages. The SSIM, UQM, PSNR, MI, and MSE are designed to compare two images. For each group of synthetic images (all six losses) we successively compare every image with each original image and then compute the average of these 40 × 40 values. We also calculate these values on all pairs of original images to compare the results to the original range. The Blur index is calculated on individual images. We store the minimum and maximum value of this index for the original images and the average value per loss for the synthetic images.

3.3. Psychovisual evaluation

3.3.1. Dataset of original and synthetic images

The dataset to evaluate contains two classes of images the 40 microscopic acquisitions mentioned in Section 3.1.1 and the



240 synthetic images created by the previous mentioned AAE loss optimizations: binary cross entropy (BCE), binary cross entropy with a L1 normalization (BCE + L1), least squares (LS), Poisson, Wasserstein (Wass.), perceptual Wasserstein (P. Wass). The size of these 280 images is 250×250 pixels.

3.3.2. Randomization

To perform a double blind test the images are not labeled during the visualization so neither the experts or the team can be biased by the images information (real, generated, nor its kind of generation). Real and generated images are randomized at each test run. Each biological expert evaluates the complete randomized dataset (280 images). The randomization and corresponding labels are stored in a .csv file which is only accessible for result analysis.

3.3.3. Experts

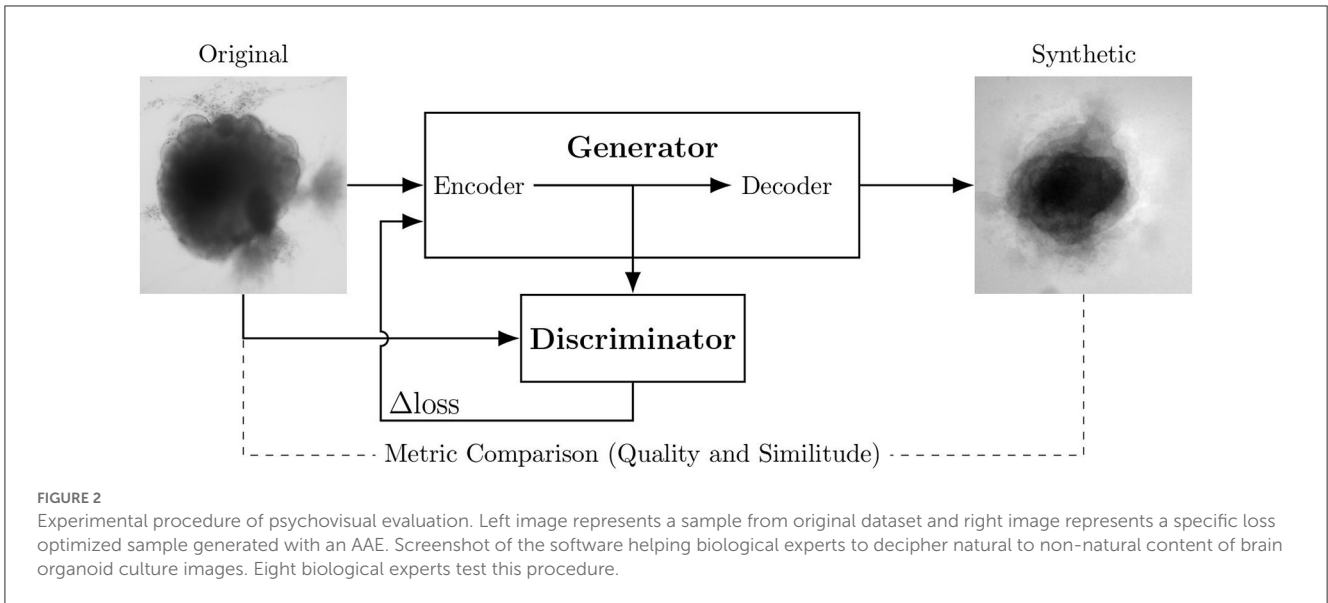
The experts who evaluated the database are biologists from ERRMECe laboratory, EA1391, CY Cergy Paris University. The group of eight experts is composed of three men and five women who are either PhD students, research engineers or researchers. They all have an expertise in neuronal culture and microscopy acquisition. We do not allow duplicate evaluators across evaluation procedure. During each evaluation session the evaluator is physically isolated from the other participants, without knowledge of other experts responses or images labels.

3.3.4. Evaluation software

To help experts in their evaluation and to ensure consistency throughout the entire experiment, we built a dedicated software using Python 3.6, as shown Figure 2. The interface consists of an image displayed (250×250 pixels), a cursor with three buttons: *Real*, *Generated*, *Next* and the number of remaining images to classify. Keyboard shortcuts are available for *Real*, *Generated*, and *Next* buttons (respectively A, P, and Tab keys on an AZERTY keyboard) to facilitate the process. Images on the screen are updated each time *Next* is hit (or the corresponding keyboard shortcut). Clicks on *Next* are counted as a pass if not preceded by a *Real* or *Generated* click.

3.3.5. Experimental protocol

An operator enters the expert name and the date and hour of the recording. All eight experts chose to use their own mouse with the experiment laptop. The protocol, consisting of a single session and including all 280 images (real and synthetic), is described in Figure 2. A pass is consider as an answer. The decision and answer time are saved at each click in a .csv file only accessible to analyze the results. The operator is present nearby to verify the smooth functioning of the experimental process and to capture any comments made by the experts. The list of questions the expert has to answer after the process is listed below:



- Why do you classify this generated image as a false? The operator shows the generated image with the longest hesitation.
- Why do you classify this original image as a false? The operator shows the original image with the longest hesitation, if this situation exists.
- What should we improve in future sessions?

We summarize the answers to these questions in the results section.

Each evaluation run produces two .csv files: One stores the randomization i.e. the order and label of each image presented. The second stores the experts' name and for each image present the answer time and the decision.

3.4. Analysis

The first analysis step consists in associating the randomization and the results files. We obtain, for each expert and for each image the decision and the decision time. The decision is then labeled as true positive (TP) or false negative (FN) for the original images and false positive (FP) or true negative (TN) for the synthetic images.

3.4.1. Parameter calculation

It is then standard to calculate the error rate (ER), defined as the number of false decisions divided by the total number of decisions.

$$ER = \frac{FP + FN}{FP + FN + TP + TN} \tag{16}$$

For the original images, this becomes:

$$ER_O = \frac{FN}{FN + TP} \tag{17}$$

and for the synthetic images:

$$ER_G = \frac{FP}{FP + TN} \tag{18}$$

However, we wish to compare the proportion of synthetic images falsely labeled as true, with the proportion of original images label as true. We thus calculate the Positive Rate (PR) of the original images:

$$PR_O = 1 - ER_O = \frac{TP}{FN + TP} \tag{19}$$

For the synthetic images $PR_G = ER_G$.

As a second parameter, we calculate the decision occurrence for each modality for each subgroup by a simple counting and render it in a % according to the total effective of a group of images.

We also evaluate the number of positive answers given by each expert as a count and the number of images given as a positive by zero expert, one expert, two experts etc. or the eight experts. Time decision and all these parameters are calculated between original and generated images, or between original and each modality of loss generation (BCE, BCE + L1, LS, Poisson, Wass., P. Wass.), globally or by each decision subgroups (FP, FN, TP, TN). All results are rendered as bargraphs representing variables (Time Decision in seconds or occurrence in % or error rates) according to one or many factors (group and subgroups of decision).

3.4.2. Metrics vs. human decision

To verify if some metrics highlight the same loss as producing the most natural images as experts, we plot each metric values for each loss group by each decision factor modality (FP, FN, TP, TN). In the dot representations for each loss group, each metric is plotted according to the Normalized Error Rate NER with individuals decision time t for the FP and for FN modality:

$$NER = \frac{FP \times t_{FP} + FN \times t_{FN}}{FP + FN + TP + TN} \tag{20}$$

where t_{FP} and t_{FN} are the average decision time respectively for FP and FN. In practice, this becomes, for the original images:

$$NER_O = \frac{FN \times t_{FN}}{FN + TP} \quad (21)$$

and for the synthetic images:

$$NER_G = \frac{FP \times t_{FP}}{FP + TN}. \quad (22)$$

To verify if a relation exists between a metric or a particular combination of metrics (M) and the decision or the time decision (DT), we calculate KL divergences on dimensional reduction results (Joyce, 2011).

$$KL = \sum_{x \in X} M_x \times \log \left(\frac{M_x}{DT_x} \right) \quad (23)$$

We only represent here kl-divergence corrpplots for each individual metric (and not metric combinations) for space considerations. We consider all possible metric combinations C :

$$C = \left\{ \binom{n}{k} \mid k \in N^*, k \leq n \right\} \quad (24)$$

where $n = 6$ is the total number of metrics. The total number of metric combinations considered is thus 63.

We then calculate Pearson and Kendall correlations between metric combinations and KL-div results (for error rate and time decision) for original and synthetic groups. We show the Pearson correlation for the ten best metric combinations. The best representation of these results (error rate or time by KL divergence) are represented as a scatter plot.

3.4.3. Statistical analysis

The normality is verified by a Shapiro and quantile to quantile graphics. We verify the homocedasticity in normal cases by a Bartlett test and in the case of non-normality by a Levene test. In case of a normality and homocedasticity, arametric tests are used (Anova), and non-parametrics tests otherwise (Kruskall-Wallis with a Tuckey *post hoc* test). Regression models are implemented to verify the interaction of factors (group and decision) on a specific variable (time or error rate or occurrence for instance). We use a *post-hoc* Holm test to compare two by two the effect of factors on variables after the regression.

We take an alpha risk at 5%. Correlation matrices are based on Pearson correlation tests.

3.5. Segmentation task

3.5.1. Dataset

We first build a training dataset composed the 40 images from the original dataset and 40 images obtained by flip-flops, rotations, whitenings, or crops of these original images. This “classical” dataset composes our baseline. We build five “psychovisual” training datasets where we replace part of the classically augmented images by synthetic images which are validated by 0, 2, 4, 6, or 8 experts.

All datasets are composed of 80 images of which 40 original but the proportion of synthetic images decreases as the number of experts required to validate an image increases. Ground truth has been manually segmented with the ITK-SNAP software (Yushkevich et al., 2006).

3.5.2. U-Net

Segmentation allows the extraction of an image content from its background. Various segmentation procedures exist but we have chosen U-Net which is widely used in the biomedical field (Ronneberger et al., 2015). U-Net has the advantage of working well for small training sets with data augmentation strategies, and has already been used for the ventricle segmentation of cleared brain organoids (Albanese et al., 2020).

3.5.3. Training

To robustly evaluate the performance of the segmentation a these small datasets we use a leave-one-out strategy where we only test on the original images. This results in 40 training sessions per dataset. We stop the training at 1,000 epochs with an average time of training of more than 1 h for each leave-one-out loop (six cases of augmentations \times 40 images = 240 h almost for the total training step). The summary of the leave-one-out strategy for every tested case is summarized in Figure 3.

3.5.4. Comparison of segmentations

To compare ground truth cerebral organoid content segmentation (GT) and U-Net (u) ones in various conditions, mean Dice scores are calculated as:

$$Dice_{(GT,u)} = \frac{2|GT \cap u|}{|GT| + |u|} \quad (25)$$

Thanks to the TP , FP , TN , and FN we could calculate the Accuracy, the Specificity, the Sensitivity, and the F1-score. The Accuracy is the ratio of true on the positives labels:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (26)$$

The Sensitivity is the ratio between how much were correctly identified as positive to how much were actually positive:

$$Sensitivity = \frac{TP}{TP + FN} \quad (27)$$

The Specificity is the ratio between how much were correctly identified as negative to how much were actually negatives:

$$Specificity = \frac{TN}{TN + FP} \quad (28)$$

The Precision is the ratio between how much were correctly identified as positives to how much were actually labeled as positives:

$$Precision = \frac{TP}{TP + FP} \quad (29)$$

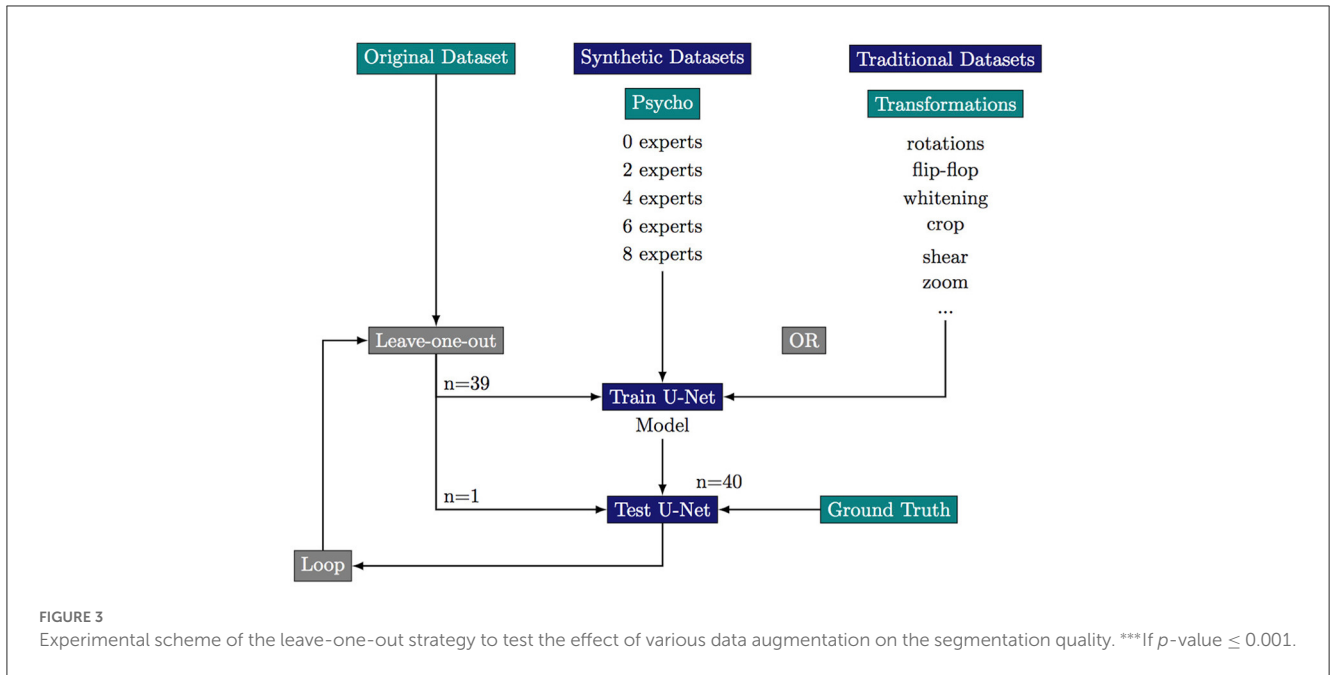


TABLE 1 Samples of original images and synthetic images generated by the AAE.

Original	BCE	BCE + L1	LS	Poisson	Wasserstein	P. Wass.

We show 3 of the 40 images for each group: original and per AAE loss variation.

The F1-Score allow to summarize the precision and the recall (Sensitivity) in an unique metric:

$$F1 - Score = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity} \quad (30)$$

3.5.5. Visualization

To highlight real/false positive/negative segmentation we create a superimposed image composed by the ground truth and a sample of each segmentation resulting from the various trainings. We update the pixels values in lightpink the *FP* cerebral organoid segmentations and, in lightgreen the *FN*.

4. Results

We first present the metric evaluation of the synthetic images, then the results of the psychovisual evaluation of these synthetic images, and finally the correlations between the metrics and the psychovisual evaluation.

4.1. Qualitative evaluation

Table 1 shows three sample input images and three of the 40 synthetic images generated by each of the six AAE variations. Some of the generated samples are blurry and present a white imprint (BCE, BCE + L1, LS). Others show sharper edges and less visible

TABLE 2 Metric evaluation of AAEGAN brain organoid bright-field generated images.

Metric	Best	Original	BCE	BCE + L1	LS	Poisson	Wass.	P. Wass.
FID	Low	0.47–0.80	1.20	1.41	1.33	1.41	1.10	0.82
SSIM	High	0.65–0.71	0.63	0.62	0.60	0.63	0.62	0.50
UQM	High	0.63–0.87	0.83	0.83	0.84	0.84	0.83	0.82
MI	High	0.21–0.47	0.37	0.39	0.36	0.41	0.46	0.42
Blur	low	0.10–86.28	135.93	116.30	135.01	106.71	59.84	59.00
PSNR	Low	11.9–16.6	13.47	13.74	13.53	13.74	13.17	12.86
MSE	Low	93.25–106.23	103.13	103.35	104.01	103.33	103.11	102.93

We have calculated metrics on generated images from each AAE loss variations, with the BCE loss as the baseline. Scores outside the original range are in gray, best values are displayed in bold.

imprints (Poisson, Wasserstein and P. Wass.). For this group of three losses, only a few of the generated images seem to be identical to a given input image. For example the Poisson loss produces three images which are a blurred version of the original. These networks do not suffer from mode collapse.

4.2. Metric evaluation

To quantitatively confirm the visual analysis of the generated images, we calculate several metrics on both the original and synthetic images. These results are summarized in Table 2. This table presents the range of values given by the original images and the average value of each group of synthetic images (per loss optimization).

All six groups of synthetic images are within range of the original for the UQM, MI, PSNR, and MSE metrics. Images generated with a Poisson or LS loss have the highest UQM index. MI and MSE reach the best scores for the generated images using the Wasserstein losses. The FID and SSIM are out of range for all six groups of synthetic images. The average FID for the Perceptual Wasserstein loss is the closest to the original (0.82 vs. 0.80). Only the Wasserstein and Perceptual Wasserstein produce images that on average are within the range of the Blur metric for the original images.

Images generated with a Wasserstein and Perceptual Wasserstein loss are on average within the range of the original images for five out of seven metrics. Quantitatively, the Wasserstein and Perceptual Wasserstein networks generate images of a quality that most reassembles the original batch. In particular, the Perceptual Wasserstein loss generates the best results for four of the seven metrics. It appears to be the most appropriate loss optimization to generate cerebral organoid images with this AAE.

4.3. Psychovisual evaluation of synthetic images

In Figure 4, we compare the occurrence of each decision in percentages for *original* and *generated* groups. There is less misleading in original and generated group than right decisions. However, 30% of misleading is observed in the generated group. A misleading corresponds to a false positive answer.

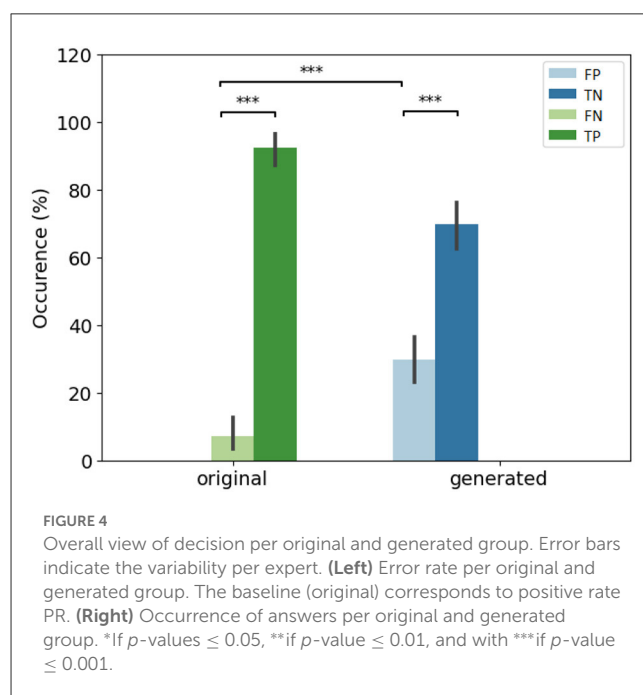
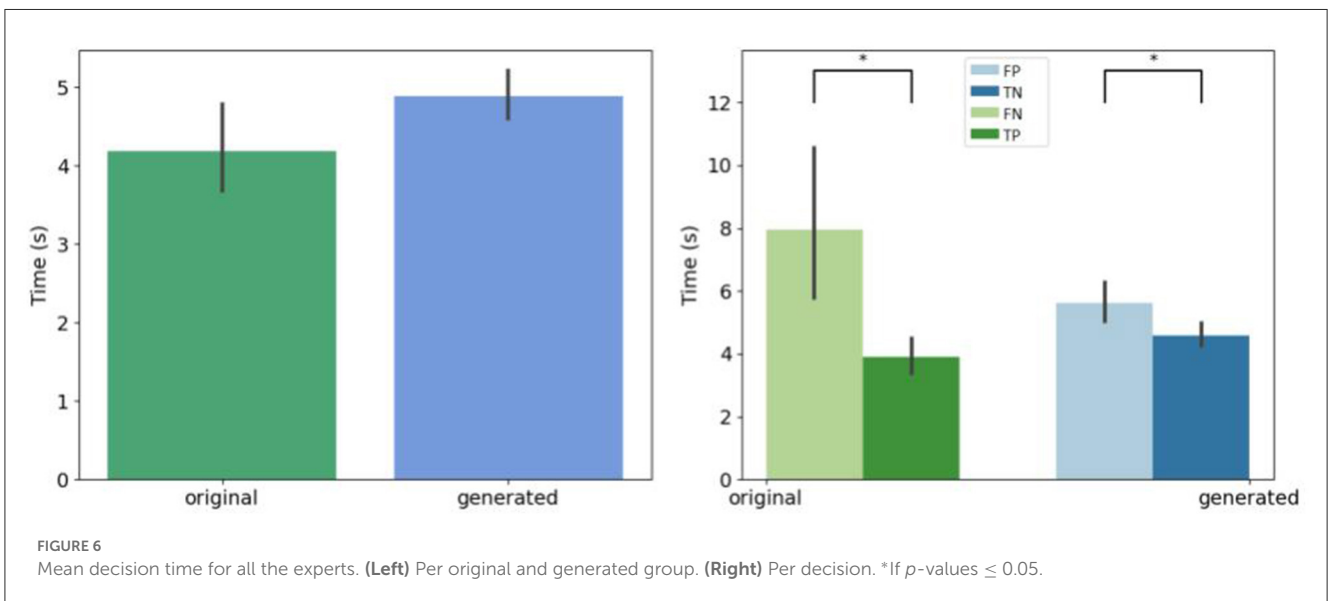
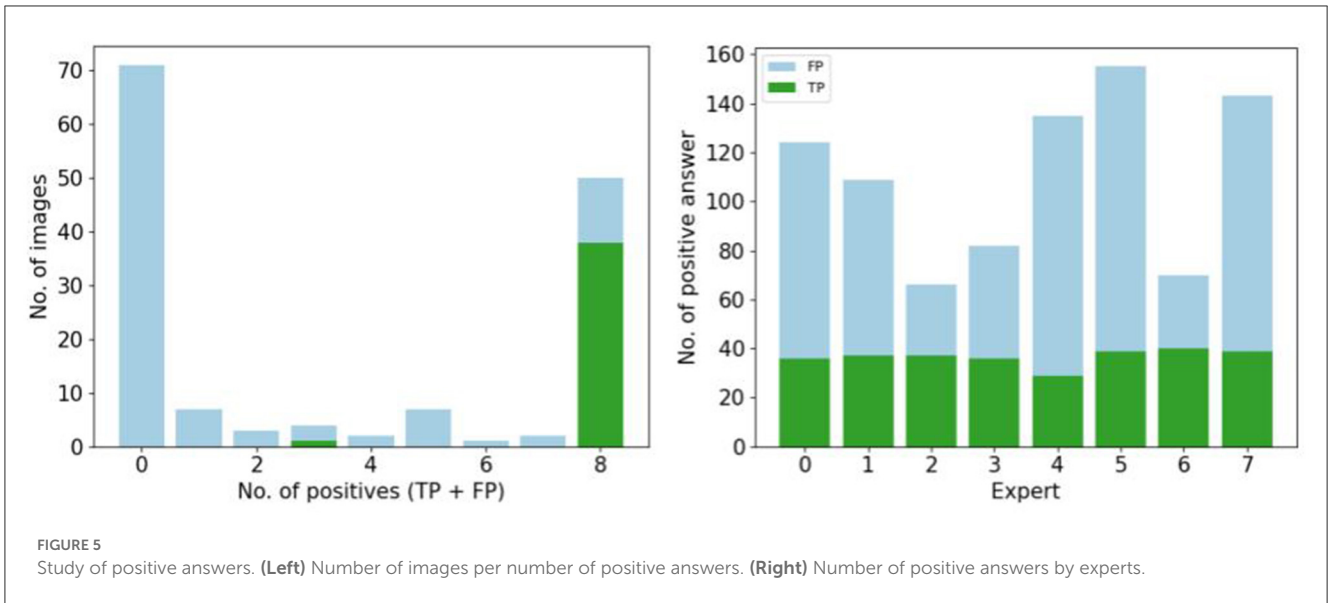


Figure 5 focuses on the images that are labeled as “natural” by the experts. We found the number of false positive selected images by all the participants is small (<20), and 30 images are selected by five participants. Almost 70 images are not selected at all by experts as natural (first column). To observe the number of false positive answer by each experts (see Figure 5). Three experts answer less positive answers than the others (less than the half of the visualized dataset). One expert considered over 150 images as natural.

We retrieve the decision time before the expert give an answer whatever the kind of generation, as shown Figure 6. Biological experts answer in the same time for generated and original images. However, the hesitation time is longer for false answers than for correct answers.

4.4. Feedback on the psychovisual procedure

Experts in cultures consider images as natural when an ovoid shape with neuroepithelial formation and some cell dispersion



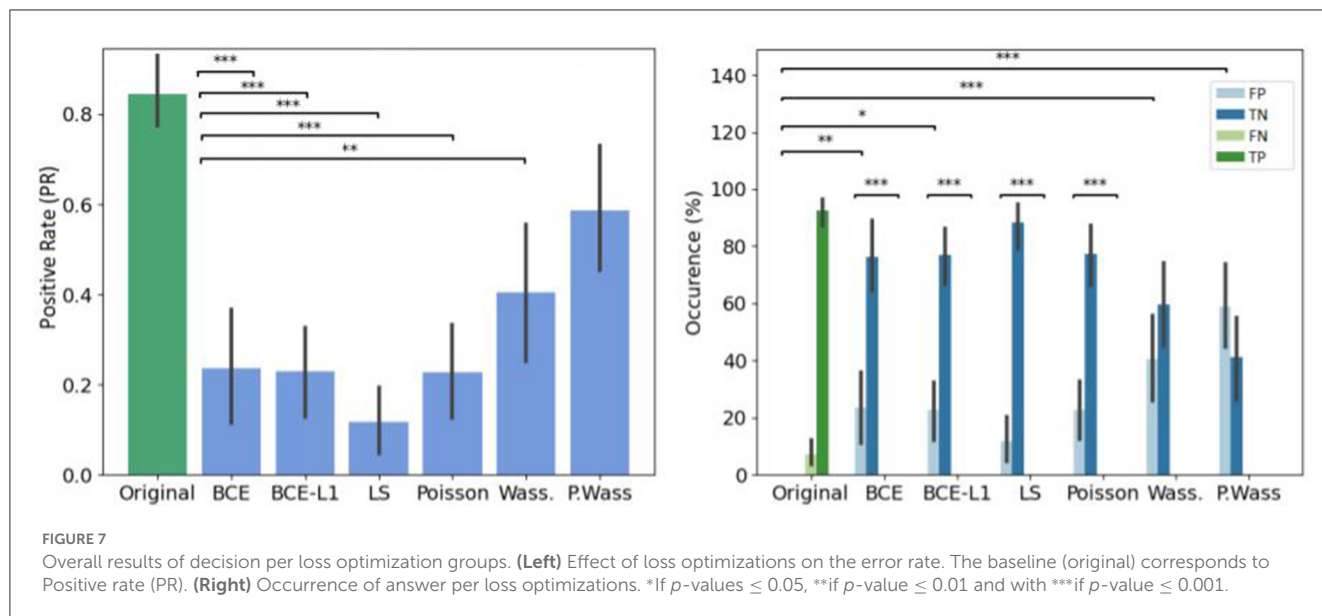
appears. When we ask experts why they classify a generated image as a synthetic, they answer the background contains an imprint, or superimposed contours or there is an artifact, or the image is too noisy, but they hesitate longer due to the possible content. They explain that they classify an original image as a synthetic because of a microscopic acquisition artifact, learned by one of the architectures, and reproduced on the worst synthetic images. They would like to have less images in a session, and a larger image on the screen.

4.5. Psychovisual evaluation of loss optimization

The error rate is particularly higher in the Wass. and P. Wass groups than for the original one (see Figure 7 left). In the P. Wass case this high score is strengthened by the absence of statistical

differences. These particular loss optimizations drive the experts to mislead and consider the images from these two groups as natural. In Figure 7 right, there is a difference between false positives of original and generated images from BCE, BCE+L1, Wass. and P. Wass. loss optimization. However, if we consider the intra-factor loss comparison, we can observe statistical differences between FP and TN of each for the BCE, BCE + L1, LS and Poisson loss rendering too small the proportion of misleading. There is no differences between these two occurrences decision for images generated by a Wass. or a P. Wass. loss showing 42% of FP and almost 60% of FP.

To observe which group of images is the most selected as positive, we observe the number of image selected by group in the Figure 8 left, P. Wass. and Wass. images are selected as natural by the most of experts (a few Poisson, and a few BCE by seven experts and L1 + BCE). The same three experts as in Figure 5 only answer positive in most of the case for P. Wass. images (see Figure 8



right). Four experts answer more synthetic images but more even for P. Wass. Only one expert seems to answer identically for all synthetic group of images.

If we do not consider the kind of decision, there is no difference of decision time per group of synthesis, (Figure 9 left). When we study the decision time per group, the experts take more time to answer only when they are confronted to synthetic images generated with a least square optimization (Figure 9 right).

4.6. Concordance of metrics and psychovisual evaluations

4.6.1. Comparisons

After observing the psychovisual decisions by generated groups, we compare qualitative and similitude metrics to the previous results in order to verify if the same groups are selected, but also to verify if some metrics or combination of metrics can be used as a proxy to human psychovisual evaluation.

An overview of these results is given in Table 3 shows no differences between decision whatever the group of loss optimization or the calculated metric. FID is the highest for *Poisson* loss than for others groups whatever the kind of decision. BLUR metric is the highest for the decision with LS generated images. In term of SSIM, UQI indexes and PSNR, the decision reach the highest score for original images and no improvement is visible with generative methods. For MSE and MI the decision rate reach the highest score similarly for original and P. Wass. generated images. No differences are visible in term of decision with UQI.

4.6.2. Correlations

To identify the metrics which best correspond to the psychovisual evaluation, we plot metrics against error rate in Figure 10. We show the Blur scatterplot as an example of point representations. In this graphic, we observe that the green color

points (*Original*) are near the darker-purple ones (P. Wass.). Figure 10 also represents the KL divergence between the P. Wass group and all other groups for all metrics. If we look at the first column, P. Wass and *Original* images are closest according to SSIM, MI and UQI. These metrics are good candidates to build a metric that mimics psychovisual evaluations.

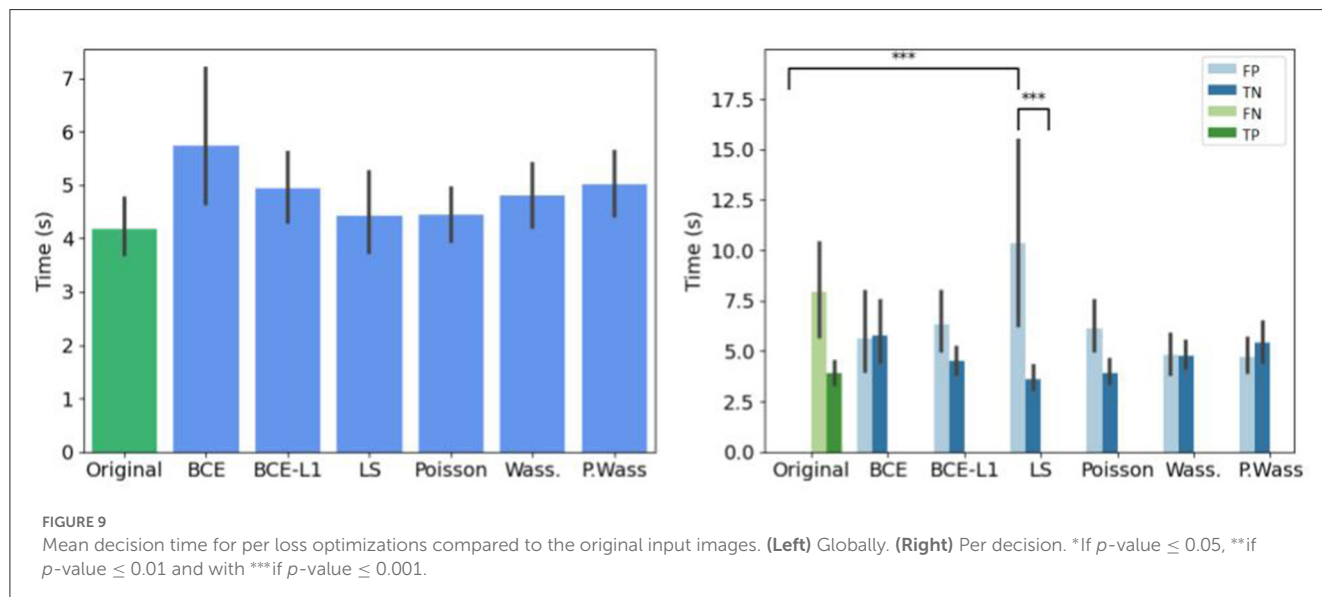
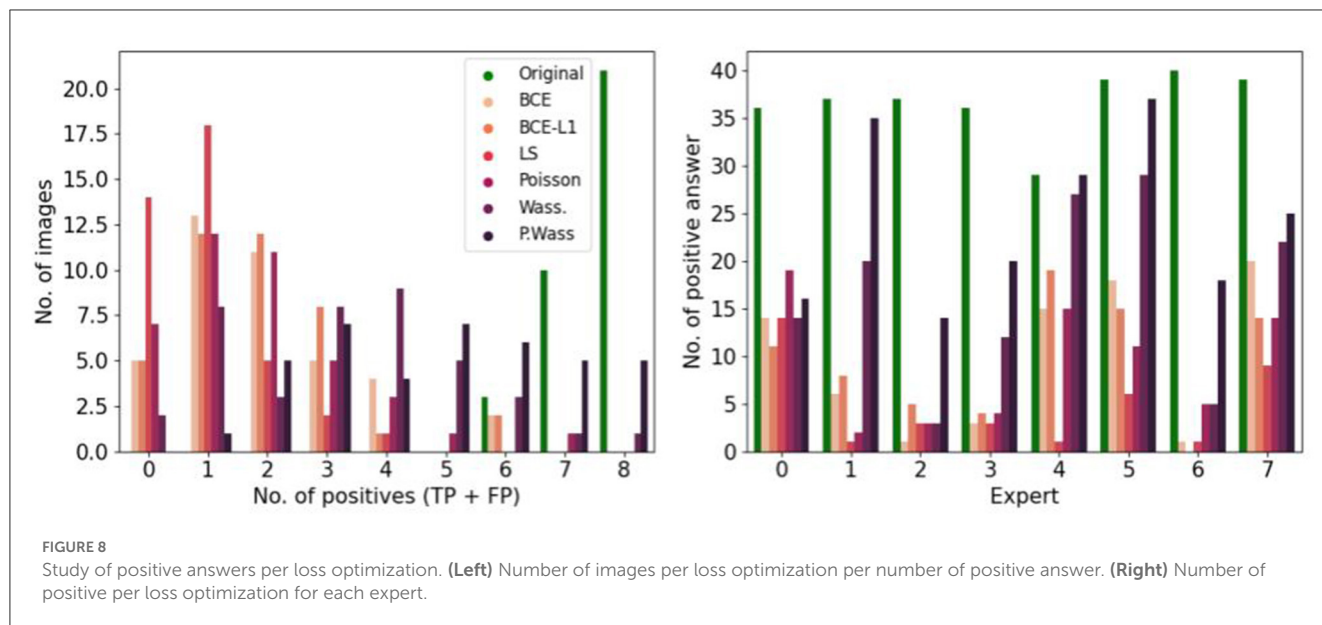
Figure 11 summarizes the correlations between the psychovisual assertion and the hesitation time or error rate according to each metric or chosen metric combinations. The main result is the absence of correlation for single metrics, however, combination with a Blur metric, FID (for the error rate correlations) and SSIM-FID-MI (for the time correlation) render the highest results in Figure 11 top left. The same result is represented in Figure 11 top right. To observe the group representation between the Error rate or the Time and the KL divergence of points represented for these two combinations (Blur-FID and Blur-SSIM-FID-MI) (see Figure 11 bottom right and left). The LS and BCE group are far from the others point representations. P. Wass. group is superimposing the original one with Wass. Others groups are not distinguishable, however, there are at the peripheral zone of the perceptual-original amount. The KL divergence representation with respect to the error rate of these two metric combinations is given in the bottom part of Figure 11.

4.7. Influence of psycho-validated images on a segmentation task

Then the second task is to verify the interest of using synthetic images which have been validated by 0, 2, 4, 6, or 8 biological experts to train a segmentation task.

4.7.1. Qualitative results

To observe the quality of the segmentation, we show a ground truth segmentation performed with the ITK-SNAP software, and



automatic segmentations performed with a U-Net architecture with various data augmentation strategies, see vignettes in Table 4. The “0 experts” group corresponds to training the segmentation with images that have been selected by none of the experts. In the others training images are previously selected by 2, 4, 6, and 8 experts. We observe less false positive regions (in pink) and false negative regions (in green) if the segmentation is performed after training on a dataset containing images validated by 6 or 8 experts. If synthetic images selected by six or more experts are used for training, we observe almost no errors on the segmentation.

4.7.2. Quantitative results

Table 4 summarizes the metrics used to compare the ground truth segmentation with the segmentation performed by U-net trained on data augmented with classic strategies or on a varying portion of synthetic images. The segmentation is better when the

network is trained with images validated by 8 experts, with higher levels of Dice, Accuracy, Sensibility and F1-score. The highest sensitivity is reached by the group trained on 41 synthetic images validated by two experts and by the group trained on images validated by no experts. However, the corresponding specificity is very low.

5. Discussion

In this part we present to our knowledge the first psychovisual and metric evaluation comparison of Loss optimized generative adversarial network of brain organoid bright-field images. This study helps at validating most natural images generated by various AAE loss optimizations. We also contribute to strengthen metric evaluation by highlighting some images from optimized generated adversarial network to be perceived by Human biological expert as natural microscopic images: with a P. Wass. loss perception

TABLE 3 Average and standard deviation of five of metrics on original and synthetic images per psychovisual decision (true or false) and per loss optimization.

	Decision	Count	Blur		SSIM		PNSR		MSE		MI		UQI	
			avg.	σ	avg.	σ	avg.	σ	avg.	σ	avg.	σ	avg.	σ
Positive														
Original	TP	297	64.15	50.28	0.78	0.17	42	41.49	2973	3163	0.90	0.46	0.86	0.12
BCE	FP	79	100.22	15.40	0.62	0.06	13.52	3.20	3713	2610	0.856	0.14	0.83	0.10
BCEL1	FP	77	104.16	38.89	0.61	0.07	13.54	2.94	3549	2320	0.82	0.13	0.83	0.09
LS	FP	39	155	44.78	0.59	0.07	13.77	3.00	3403	2276	0.78	0.11	0.84	0.09
Poisson	FP	74	93.17	22.76	0.63	0.06	13.86	2.84	3275	2139	0.86	0.14	0.84	0.09
Wass.	FP	134	61.47	18.25	0.61	0.06	13.33	2.38	3499	1980	0.90	0.14	0.83	0.07
P. Wass.	FP	197	47.19	17.13	0.63	0.07	12.48	2.44	4272	2389	0.95	0.17	0.80	0.07
Negative														
Original	FN	23	44.93	25.36	0.71	0.01	10.10	0.90	6418	1326	0.88	0.23	0.78	0.04
BCE	TN	241	105.01	20.95	0.62	0.06	13.57	3.16	3647	2555	0.85	0.13	0.83	0.10
BCEL1	TN	243	110.43	39.30	0.60	0.07	13.60	2.87	3477	2248	0.81	0.13	0.83	0.09
LS	TN	281	156	44.21	0.58	0.06	13.79	2.80	3307	2118	0.78	0.11	0.84	0.09
Poisson	TN	246	106.12	25.34	0.62	0.06	13.86	2.77	3248	2090	0.84	0.13	0.84	0.09
Wass.	TN	186	56.96	17.18	0.63	0.07	13.41	2.58	3510	2116	0.90	0.13	0.83	0.07
P. Wass.	TN	123	41.22	14.04	0.63	0.07	12.69	2.55	4123	2414	0.95	0.16	0.80	0.075

Best values are displayed in bold.

neurobiologists are misled 60% of the time and 40% by Wass. They take more time to answer when they are misled. We compare human and metric evaluation and found mutual information to be the most related to their decision metric, although no correlation appeared in our experiment for single metrics, but only for combinations including blur. Using synthetic images validated by an increasing number of experts to train a segmentation network increases the accuracy of the segmentation with respect to classic augmentation strategies, even if the proportion of synthetic images decreases.

Synthetic images generated with AAE are coherent with original dataset and thus whatever the kind of loss optimization. The generation of brain organoid images with others architectures does not improve the synthesis in term of quality or similitude according to Brémond Martin et al. (2022a), whereas it seems the case with loss optimization. The P. Wass. loss optimization of AAE performs best according to metrics. Other loss optimizations show also high similitude, though with a lower quality. In this context, we plan to explore what type of information each loss brings during the image generation. We aim at trying others embedded losses (already used for segmentation tasks) during the generative process based upon high level prior like object shape, size topology or inter-regions constraints (El Jurdi et al., 2021). These losses could be used on condition that the morphological development of CO is better characterized.

Biomedical experts select around 40% of synthetic images as natural compared to the original dataset. Thus, the generation by AAE networks generate a large part of realistic images such as the background of bright-field acquisition or, their content. The

non-selected images where considered sometimes as non-natural due to some artifacts reproduced in some of them, or by a superposition of contours. Nevertheless, the selected images can help train a DL segmentation network.

A first argument of the strong validation of the selected images as natural is the time to take a decision (Shaffrey et al., 2002). If the time to answer natural for a generated group corresponds to the time to answer natural or original images, we could consider these two groups are perceived as similar. We found no differences between original and generated images and thus whatever the kind of loss optimization used to produce them. So they are not doubting when they classify an image as natural or generated. However psychovisual evaluation shows an increase of decision time before answering when they answer as false positive (depending of the loss optimization) or false negative. This behavior is specifically shown from a Least Square Loss Optimization generated images considered as natural. When we ask participants why they have doubt on a particular image, they answer that it was linked with some acquisition artifact learned by the generated process and found on a lot of images (a bunch of cells) or, by a blurry contour which could be due to the acquisition in the case of original images (Ali et al., 2022). For false negative answers, they only said that the artifact acquisition is also present (and they thought it was a generated). In the future, we think a pre-process image treatment has to be done on images to correct the acquisition artifact before the generative process, to avoid these false negative in the psychovisual evaluation or, to add a component in the generative network to avoid these artifacts (Galteri et al., 2017; Ali et al., 2022).

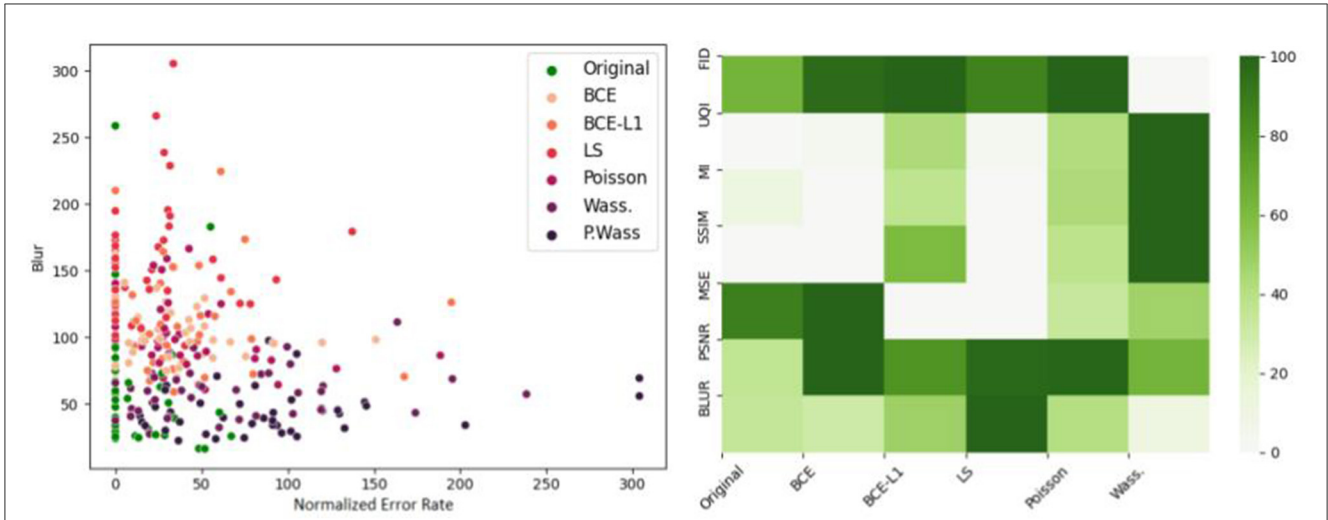


FIGURE 10 Comparison of error rate per optimization losses and per metrics. (Left) Blur score with respect to the normalized error rate for all 280 images grouped by optimization loss. (Right) KL divergence between the P. Wass group and all other groups for all metrics. Dark green represents a strong divergence between. Groups with a small KL divergence groups with respect to P. Wass. are white.

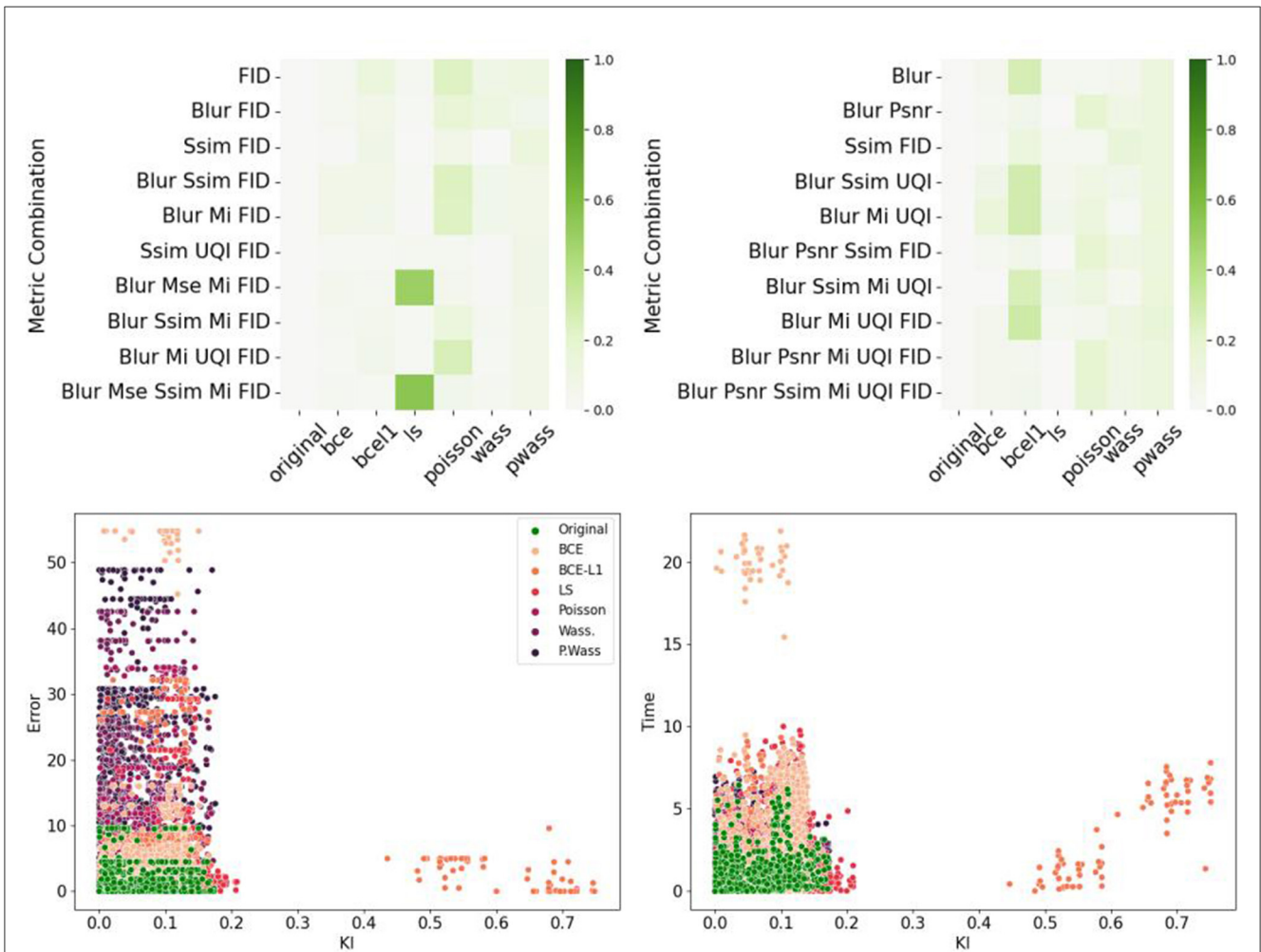


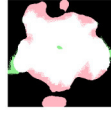
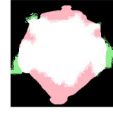





FIGURE 11 Correlations between metrics and psychovisual assessment on all 240 synthetic images. (Left) Error rate. (Right) Hesitation time. (Top) Correlation matrix of the ten best combinations. (Bottom) Example of error rate (resp. hesitation time) over KL divergence for a given metric combination: (Left) Blur and FID. (Right) Blur SSIM MI and UQI.

TABLE 4 Segmentation results after training on a varying number of synthetic images.

	GT	Classical	No. of experts validating the synthetic images				
			0	2	4	6	8
No. of synthetic images	–	0	33	41	22	16	14
Sample result							
Dice	1.0	0.80	0.62	0.59	0.61	0.67	0.82
Accuracy	1.0	0.77	0.63	0.61	0.68	0.62	0.93
Sensitivity	1.0	0.94	0.96	0.97	0.87	0.83	0.91
Specificity	1.0	0.92	0.50	0.48	0.60	0.56	0.94
F1-score	1.0	0.84	0.64	0.61	0.64	0.59	0.87

Seventy-nine images are used for training of which 39 are from the original data set. The remaining 40 are a combination of classic transformations and synthetic images. Tests performed with no synthetic images (classical) are compared to those performed with synthetic images validated with an increasing number of experts (0 to 8). Best values per metric are displayed in bold.

Only 15 images are selected by all the experts as natural. Five experts select over 100 images as natural and three <80. This study raises a question: could we use images considered by only five experts as natural in a training step? Thus we thought we need to increase the number of biological experts to overcome future studies in order to be more precise on the number of images considered as natural. We could also analyze the answers by the field of expertise of biological experts too (separate those who made only culture or only microscopic acquisition from those working in both fields).

Nonetheless, human Psychovisual experts choose in majority images from generative adversarial network as natural if they are from Wass. and P. Wass. loss and, a few BCE, BCE-L1. These two first kinds of generation are also highlighted by most of the metrics in an other study to have the better quality and to be the most similar to original images (Brémond Martin et al., 2022a). Thus, the psychovisual evaluation strengthen the choice of the use of these two and particularly the perceptual one in generative process. We can now confirm the idea that the regulation term of the Wass. distance between two images (Kupyn et al., 2018) could improve the learning of the pattern or characteristics of brain organoids in images and contribute to generate more natural images in term of content and aspects. In future studies we will remove the images that did not dupe the experts to only train on human-validated images. We would like to see if this increases the segmentation accuracy and study the impact on the morphological characterization.

However, the few BCE and BCE-L1 images selected as natural by psychovisual experts could maybe have also a great interest whereas the metric are not pointing them as natural images (Brémond Martin et al., 2022a). As we know the use of metric is still controversial for the GAN evaluation as they are measuring similitude and quality (Borji, 2019). Here, we could not highlight a strong correlation between the use of certain metrics and the decision to reject or not a generated image as natural. To correlate a metric and psychovisual evaluation instead of a binary answer “natural” or “not natural,” some authors use a graduation scale

(Pedersen and Hardeberg, 2012; Pedersen, 2015). This approach could be tested in future studies.

No metric used in this study could replace a human perceptual evaluation to decipher the naturality of an image generated. There is a certain link with FID, BLUR or MI and the group and MI with the mean decision but it remains weak. We could only say that similitude and referenced-bases metrics are more linked to the decision than qualitative metrics and non-reference-based metrics. And when we compare metrics with decisions some patterns appear according to the kind of loss optimization. The use or not of a measure to decipher natural generated examples is an issue recently discussed (Borji, 2019). To compare fairly images generated by various optimized models, there is no consensus for a use of a particular metric. In other fields such metric comparisons highlight a wavelet structural similitude index WSSI, a metric which based upon SSIM but less complex and more accurate in term of quality assessment (Rezazadeh and Coulombe, 2009; Pedersen, 2015). However, we do not want an identical image but one just resembling as a natural one. This could explain these metrics are not well designed for the GAN specific evaluation when they are considered alone. This study comparing the overall psycho-visual evaluation and seven metrics is one of the pioneer work which could contribute to help at pointing a metric of “natural,” and it failed partially.

Based upon our KL divergence maps, we suggest that a combination of metrics which best represents the psychovisual evaluation decision (BLUR, SSIM, MI, UQI) could be used as a substitute for a human psychovisual evaluation which is time consuming. Nevertheless, this work on metric combinations replacing a psychovisual evaluations need to be further studied. In other fields the combination of metrics help at pointing out some results in term of quality or similitude (Yao et al., 2005; Pedersen and Hardeberg, 2012; Okarma et al., 2021). An other idea could also to use non-reference quality metrics combinations (Rubel et al., 2022). Some authors tries also to implement directly a discriminator of generative adversarial networks based upon human perception, this could be a solution if it is not time consuming (Fujii et al., 2020; Arnout et al., 2021). It is not the case

in this study, for us, an important task is to find an appropriate metric for highlighting “the naturality” of the image and replacing the psycho-visual evaluation. An idea is to test psycho-metrics instead of classical similitude or quality metrics such as Hype from Zhou et al. (2019) which is an alternative of *FID* from Heusel et al. (2017), or implementing the GFI quality assessment created by Tian et al. (2022).

The more experts validate a portion of synthetic images in the dataset, the better the segmentation quality. This suggests that if more experts are available to select images, and strengthen the naturality of the synthetic dataset used during the training, it could improve the accuracy of the segmentation results. However, even if the psychovalidation of certain synthetic images allows us to improve the segmentation, this method is still subjective. It could include biases of the experts about the model, its configuration, and the project objective. It requires knowledge of which images are considered as natural and which ones are not for the target domain, so the number of experts available in the field diminished while our task required more experts. It is limited to the number of images that can be reviewed in a reasonable time (Booij et al., 2019).

Validation by six experts is the minimum to improve the segmentation, but quantitative analysis show us, eight experts validation is the minimum due to the equilibrium state between the specificity and the sensibility. The performance of human judges is not fixed and can improve over time, other articles choose a validation by 15 experts for instance which is not possible in our biomedical context which requires experts in the field (Denton et al., 2015; Salimans et al., 2016b).

We could also apply this psychovisual evaluation on others datasets to attempt to answer more specifically to the metric replacement. We thought about noise optimized generated images of brain organoids with an AAE for the same aim (Brémond Martin et al., 2022b). It could be interesting to observe if with a noise injection, similar to the bright-field acquisition images, generated images are more perceive as natural even if metrics are not pointing a particular kind of noise. Indeed, qualitative and similitude metrics point out Gaussian noise and shot noise injection. But as said previously, this could maybe only due to the metric choice (Borji, 2019). An analysis of Psychovisual evaluation could maybe help at highlighting a combination of metrics. In this future study, it could be also interesting to observe for example the microscopic experience of the Biological expert as a new criterion. A larger application of this methodology could be made on others kind of generation (such as on GAN, Goodfellow et al., 2014 or DCGAN, Radford et al., 2015) and maybe help at pointing out the best GAN model for brain organoid generation used during a training segmentation task.

In this study we use a unbalanced dataset with more synthetic images than original. Nevertheless, biological experts do not know the number of real or synthetic images which render it unbiased. In future studies, we need to obtain and use more original images in order to re-equilibrate. We use a software created specifically for the psychovisual task for brain organoid images. The software needs to be updated due to some limitations. We have to realize batch process with pauses to limit the tiredness of biological experts similar to others psychovisual evaluations (Shaffrey et al., 2002). We

have to add also a cursor with a score instead of a button to estimate a natural range in future studies and facilitate correlations studies (Tian et al., 2022). The size of the image of the screen has to be increase but not for all the participants. Apart from these updates, the use of the software is simple and practical according to their feedback.

To strengthen our statistical analysis we should increase the number of biological experts. However, it could include biases of the experts: it requires knowledge of which image is considered as natural and which one is not for the target domain, so the number of experts available in the field is diminished while our task required more experts. The performance of human judges is not fixed and can improve over time, other articles choose a validation by 15 experts for instance which is not possible in our biomedical context which requires experts in the field (Denton et al., 2015; Salimans et al., 2016a). Moreover, psychovisual evaluation is limited to the number of images that can be reviewed in a reasonable time (Borji, 2019). The tradeoff between the number of synthetic images used to train a network and the number of validating experts could be further explored.

6. Conclusion

In this study psycho-visual evaluations allow us to:

- Validate some synthetic image generated from loss optimization of generative brain organoid images with an AAE in term of decision time and decision.
- Describe the quality and similitude of the synthetic images with the original dataset by a metric validation.
- Verify if some synthetic images could be considered as natural by psychovisual expert decision.
- Compare psychovisual and metric evaluations.
- Paves the way to finding a metric or a metric combination that mimics psychovisual evaluations.
- Show the interest of selecting images validated by the highest number of experts in a data augmentation strategy for a segmentation task.

This selected images could be use in the training phase of a segmentation task in order to help at their morphological development characterization for instance. We also need to evaluate psychovisually noise injected optimized synthesized images.

In future studies we suggest a combination of metrics or a perceptual metric could maybe help at replacing the psycho-visual assessment which is time consuming. Such methodology could be used for others brain organoid data-sets generated with a generative adversarial network.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

This article was an idea of CB-M, CS-C, CC, and AH. The literature search and data analysis was performed by CB-M. Graphics tables and figures were conceived by CB-M and CS-C. The first draft was written by CB-M while CS-C, CC, and AH critically revised the work. All authors contributed to the article and approved the submitted version.

Funding

ANRT supported CB-M Ph.D. scholarship (2019/0635).

Acknowledgments

The authors would like to thank Dr. Michel Boissière for his assistance and feedback. The authors would like to acknowledge all

References

- Albanese, A., Swaney, J. M., Yun, D. H., Evans, N. B., Antonucci, J. M., Velasco, S., et al. (2020). Multiscale 3D phenotyping of human cerebral organoids. *Sci. Rep.* 10, 21487. doi: 10.1038/s41598-020-78130-7
- Ali, M. A. S., Hollo, K., Laasfeld, T., Torp, J., Tahk, M.-J., Rinken, A., et al. (2022). AttSeg—artifact segmentation and removal in brightfield cell microscopy images without manual pixel-level annotations. *Sci. Rep.* 12, 11404. doi: 10.1038/s41598-022-14703-y
- Alqahtani, H., Kavakli, M., and Kumar Ahuja, D. G. (2019). “An analysis of evaluation metrics of gans,” in *International Conference on Information Technology and Applications*, Vol. 7 (Sydney, NSW).
- Arnout, H., Bronner, J., and Runkler, T. (2021). “Evaluation of generative adversarial networks for time series data,” in *2021 International Joint Conference on Neural Networks* (Shenzhen: IEEE), 1–7. doi: 10.1109/IJCNN52387.2021.9534373
- Balaji, Y., Chellappa, R., and Feizi, S. (2019). “Normalized wasserstein for mixture distributions with applications in adversarial learning and domain adaptation,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul: IEEE), 6499–6507. doi: 10.1109/ICCV.2019.00660
- Booij, T. H., Price, L. S., and Danen, E. H. J. (2019). 3D Cell-based assays for drug screens: challenges in imaging, image analysis, and high-content analysis. *SLAS Discov.* 24, 615–627. doi: 10.1177/2472555219830087
- Borji, A. (2019). Pros and cons of gan evaluation measures. *Comput. Vis. Image Underst.* 79, 41–65. doi: 10.1016/j.cviu.2018.10.009
- Brémond Martin, C., Simon Chane, C., Clouchoux, C., and Histace, A. (2021). Recent trends and perspectives in cerebral organoids imaging and analysis. *Front. Neurosci.* 15, 629067. doi: 10.3389/fnins.2021.629067
- Brémond Martin, C., Simon Chane, C., Clouchoux, C., and Histace, A. (2022a). “AAEGAN loss optimizations supporting data augmentation on cerebral organoid bright field images,” in *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Volume 4, VISIGRAPP* (Setúbal: SciTePress), 307–314. doi: 10.5220/0010780000003124
- Brémond Martin, C., Simon Chane, C., Clouchoux, C., and Histace, A. (2022b). “AAEGAN optimization by purposeful noise injection for the generation of bright-field brain organoid images,” in *Internal Conference On Image Processing Theory Tools and Application: Special Session Biological and Medical Image Analysis* (Salzburg: IEEE), 6. doi: 10.1109/IPTA54936.2022.9784149
- Chakradhar, S. (2016). New company aims to broaden researchers’ access to organoids. *Nat. Med.* 22, 338–338. doi: 10.1038/nm0416-338
- Chen, Y., Yang, X.-H., Wei, Z., Heidari, A. A., Zheng, N., Li, Z., et al. (2022). Generative Adversarial Networks in Medical Image augmentation: a review. *Comput. Biol. Med.* 144, 105382. doi: 10.1016/j.compbiomed.2022.105382
- Denton, E., Chintala, S., Szlam, A., and Fergus, R. (2015). “Deep generative image models using a laplacian pyramid of adversarial networks,” in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS’15* (Cambridge, MA: MIT Press), 1486–1494.
- El Jurdi, R., Petitjean, C., Honeine, P., Cheplygina, V., and Abdallah, F. (2021). High-level prior-based loss functions for medical image segmentation: a survey. *Comput. Vis. Image Underst.* 210, 103248. doi: 10.1016/j.cviu.2021.103248
- Fernandez, R., Rosado, P., Vegas, E., and Reverter, F. (2021). Medical image editing in the latent space of generative adversarial networks. *Intell.-Based Med.* 5, 100040. doi: 10.1016/j.ibmed.2021.100040
- Fujii, K., Saito, Y., Takamichi, S., Baba, Y., and Saruwatari, H. (2020). “HUMANGAN: generative adversarial network with human-based discriminator and its evaluation in speech perception modeling,” in *IEEE International Conference on Acoustics, Speech and Signal Processing* (Barcelona: IEEE), 6239–6243. doi: 10.1109/ICASSP40776.2020.9053844
- Galteri, L., Seidenari, L., Bertini, M., and Bimbo, A. D. (2017). “Deep generative adversarial compression artifact removal,” in *IEEE International Conference on Computer Vision* (Venice: IEEE), 4836–4845. doi: 10.1109/ICCV.2017.517
- Gomez-Giro, G., Arias-Fuenzalida, J., Jarazo, J., Zeuschner, D., Ali, M., Possemis, N., et al. (2019). Synapse alterations precede neuronal damage and storage pathology in a human cerebral organoid model of CLN3-juvenile neuronal ceroid lipofuscinosis. *Acta Neuropathol. Commun.* 7, 222. doi: 10.1186/s40478-019-0871-7
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial networks. *Adv. Neural Inf. Process. Syst.* 3.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017). “Improved training of wasserstein gans,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17* (Red Hook, NY), 5769–5779.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17* (Red Hook, NY: Curran Associates Inc), 6629–6640.
- Hu, B., Tang, Y., Chang, E. I.-C., Fan, Y., Lai, M., Xu, Y., et al. (2019). Unsupervised learning for cell-level visual representation in histopathology images with generative adversarial networks. *IEEE J. Biomed. Health Inf.* 23, 1316–1328. doi: 10.1109/JBHI.2018.2852639
- Jiang, M., Zhi, M., Wei, L., Yang, X., Zhang, J., Li, Y., et al. (2021). Fa-gan: fused attentive generative adversarial networks for mri image super-resolution. *Comput. Med. Imaging Graph.* 92, 101969. doi: 10.1016/j.compmedimag.2021.101969
- Joyce, J. M. (2011). “Kullback-Leibler divergence,” in *International Encyclopedia of Statistical Science*, ed M. Lovric (Berlin: Springer Berlin Heidelberg), 720–722. doi: 10.1007/978-3-642-04898-2_327
- the team from ERRMECe laboratory for participating to blinded psychovisual evaluations of synthetic images.

Conflict of interest

CB-M and CC were employed by Witsee, Neoxia.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Kelava, I., and Lancaster, M. A. (2016). Dishing out mini-brains: current progress and future prospects in brain organoid research. *Dev. Biol.* 420, 199–209. doi: 10.1016/j.ydbio.2016.06.037
- Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., and Matas, J. (2018). “DeblurGAN: blind motion deblurring using conditional adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 8183–8192. doi: 10.1109/CVPR.2018.00854
- Lan, L., You, L., Zhang, Z., Fan, Z., Zhao, W., Zeng, N., et al. (2020). Generative adversarial networks and its applications in biomedical informatics. *Front. Public Health* 8, 164. doi: 10.3389/fpubh.2020.00164
- Lancaster, M. A., Renner, M., Martin, C.-A., Wenzel, D., Bicknell, L. S., Hurler, M. E., et al. (2013). Cerebral organoids model human brain development and microcephaly. *Nature* 501, 373–379. doi: 10.1038/nature12517
- Mahmood, F., Borders, D., Chen, R. J., McKay, G. N., Salimian, K. J., Baras, A., et al. (2018). Adversarial training for multi-organ nuclei segmentation in computational pathology images. *IEEE Trans. Med. Imaging* 39, 3257–3267. doi: 10.1109/TMI.2019.2927182
- Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. (2016). Adversarial autoencoders. *arXiv*. [preprint]. doi: 10.48550/arXiv.1511.05644
- Malm, P., Brun, A., and Bengtsson, E. (2015). Simulation of brightfield microscopy images depicting papsmear specimen. *Cytometry Part A* 87, 212–226. doi: 10.1002/cyto.a.22624
- Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., Smolley, S. P., et al. (2017). “Least squares generative adversarial networks,” in *IEEE International Conference on Computer Vision* (Venice: IEEE), 2813–2821. doi: 10.1109/ICCV.2017.304
- Okarma, K., Lech, P., and Lukin, V. V. (2021). Combined full-reference image quality metrics for objective assessment of multiply distorted images. *Electronics* 10, 2256. doi: 10.3390/electronics10182256
- Pedersen, M. (2015). “Evaluation of 60 full-reference image quality metrics on the CID:IQ,” in *IEEE International Conference on Image Processing* (Quebec City, QC: IEEE), 5. doi: 10.1109/ICIP.2015.7351068
- Pedersen, M., and Hardeberg, J. (2012). Full-reference image quality metrics: classification and evaluation. *Found. Trends Comput. Graph. Vis.* 7, 1–80. doi: 10.1561/06000000037
- Plum, J., Maintz, J., and Viergever, M. (2003). Mutual-information-based registration of medical images: a survey. *IEEE Trans. Med. Imaging* 22, 986–1004. doi: 10.1109/TMI.2003.815867
- Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv*. [preprint]. doi: 10.48550/arXiv.1511.06434
- Rezagadeh, S., and Coulombe, S. (2009). “A novel approach for computing and pooling structural similarity index in the discrete wavelet domain,” in *2009 16th IEEE International Conference on Image Processing (ICIP)* (Cairo: IEEE), 2209–2212. doi: 10.1109/ICIP.2009.5413886
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. *arXiv*. [preprint]. doi: 10.48550/arXiv.1505.04597
- Rubel, A., Ieremeiev, O., Lukin, V., Fastowicz, J., and Okarma, K. (2022). Combined no-reference image quality metrics for visual quality assessment optimized for remote sensing images. *Appl. Sci.* 12, 1986. doi: 10.3390/app12041986
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., et al. (2016a). “Improved techniques for training gans,” in *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS’16* (Red Hook, NY: Curran Associates Inc), 2234–2242.
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., et al. (2016b). Improved techniques for training gans. *Adv. Neural Inf. Process. Syst.* 29.
- Shaffrey, C. W., Jermyn, I. H., and Kingsbury, N. G. (2002). Psychovisual evaluation of image segmentation algorithms. *Adv. Concepts Intell. Vis. Syst* 7.
- Singh, N. K., and Raza, K. (2021). “Medical image generation using generative adversarial networks,” in *Health Informatics: A Computational Perspective in Healthcare*, eds R. Patgiri, Biswas A., and P. Roy (Singapore: Springer Nature Singapore), 19. doi: 10.1007/978-981-15-9735-0_5
- Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., et al. (2016). Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans. Med. Imaging* 35, 1299–1312. doi: 10.1109/TMI.2016.2535302
- Tian, Y., Ni, Z., Chen, B., Wang, S., Wang, H., Kwong, S., et al. (2022). Generalized visual quality assessment of gan-generated face images. *arXiv*. [preprint]. doi: 10.48550/arXiv.2201.11975
- Tsomo, E., Kim, H. J., Paik, J., and Yeo, I.-K. (2008). Efficient method of detecting blurry images. *Journal of Ubiquitous Convergence Technology*. 2, 27.
- Wang, Z., and Bovik, A. (2002). A universal image quality index. *IEEE Signal Process. Lett.* 9, 81–84. doi: 10.1109/97.995823
- Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861
- Wargnier-Dauchelle, V., Simon-Chane, C., and Histace, A. (2019). Retinal blood vessels segmentation: improving state-of-the-art deep methods. computer analysis of images and patterns. CAIP 2019. *Commun. Comput. Inf. Sci.* 1089, 5–16. doi: 10.1007/978-3-030-29930-9_1
- Yao, S., Lin, W., Ong, E., and Lu, Z. (2005). “Contrast signal-to-noise ratio for image quality assessment,” in *IEEE International Conference on Image Processing 2005* (Genova: IEEE), I-397. doi: 10.1109/ICIP.2005.1529771
- Yi, X., Walia, E., and Babyn, P. (2019). Generative adversarial network in medical imaging: a review. *Med. Image Anal.* 58, 101552. doi: 10.1016/j.media.2019.101552
- Yushkevich, P. A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J. C., et al. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* 31, 1116–1128. doi: 10.1016/j.neuroimage.2006.01.015
- Zhao, H., Gallo, O., Frosio, I., and Kautz, J. (2017). Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imaging* 3, 47–57. doi: 10.1109/TCI.2016.2644865
- Zhou, S., Gordon, M. L., Krishna, R., Narcomey, A., Morina, D., Bernstein, M. S., et al. (2019). Hype: human eye perceptual evaluation of generative models. *LCTR*, 15.