CLEANSE: controlling the quality and diversity of synthetic data characteristics for Machine Learning training

Thomas Gonzalez[1] and Rémi Pétiot[1]

[1] *Oktal-Synthetic Environment, 11 avenue du Lac, 31320 Vigoulet-Auzil, France*

December 2025

Abstract

SE-WORKBENCH (SE-WB)-EO, also called CHORALE, is a comprehensive set of tools that aims to model the $3D$ environment with a virtual and synthetic approach and to render the $3D$ scene for a given Infrared (IR) or visible sensor. The SE-WB-EO tools enable users to calculate spectral rendering data in a wide range of fields, such as infrared.

The importance of synthetic data has never been greater, and even more so since their usefulness for Machine Learning (ML) applications. For example, Deep Learning (DL) algorithms require large datasets for their training, which are intended to be as exhaustive and realistic as possible.

However, when it comes to Artificial Intelligence (AI) algorithms for synthetic data analysis, there are many unknowns. For example, SE-WB-EO is capable of generating synthetic images in Electro-Optical (EO) domains such as IR from Short Wave Infrared (SWIR) to Long Wave Infrared (LWIR). Indeed, the observables are different and the features extracted by the ML algorithms should therefore be different.

These observables are intrinsically linked to material physics. The main drawback of synthetic images is their entropy, which is not yet consistent with ground truth. In SE-WB-EO, a colour, in the visible spectrum, is linked to a physical material, which is consistent with the SWIR to LWIR domains. Currently, the limitation today is the lack of colours in the visible domain for synthetic spectral visual images.

However, when it comes to ML algorithms, synthetic data, even if it differs from real data in terms of its lack of colours, for example, remain useful in cases where there is a shortage of data, also called data scarcity. Nevertheless, it is important to be able to measure the quality of a dataset intended to feed a ML algorithm for learning purposes. The problem is even more significant when, as with the SE-WB, there is no real data available to compare the discrepancy with reality. To this end, in this study we present the first version of a prototype whose objective is to provide indicators on the diversity and quality of a dataset intended for use in ML.

The first part presents a state of art about synthetic data for DL training and the current state of SE-WB-EO. The reality gap of synthetic simulators is also mentioned, particularly with regard to visual spectral rendering.

Next, the new approach is presented in details.

Finally, before concluding, results are shown and we mentioned the limitation of the application.

The next step to be taken in the future will be presented.

Keywords: CHORALE, CLEANSE, Infrared, Machine Learning, Metrics, Pixels, RGB, Robustness, SE-WORKBENCH, Simulation, Spectral rendering, Synthetic data.

1 Introduction

Recently, an alternative to SE-WB, a Radio-Frequency (RF) and EO simulation toolkit, has been developed. The aim of SE-WORKBENCH-AI (SE-WB-AI) is to provide a solution to the data scarcity through a massive data generator. Data scarcity is a significant problem in the field of ML due to the importance of data in the learning stage of a DL algorithm. Data is the cornerstone of the inference performance of ML algorithms. However, for some tasks, data is difficult to obtain and, in some cases, unavailable due to its sensitivity. In most cases, it is quite common for datasets to be unlabelled, requiring considerable effort to label them. Furthermore, the labelling process is difficult and prone to errors, posing a considerable risk to the training of ML algorithms and making bias avoidance more complex. Furthermore, some data, particularly image quality, are highly dependent on environmental conditions at the time of acquisition. For target detection, for example, this means that the data acquired must be correlated with atmospheric conditions such as weather, solar radiation, temperature, and pollution.

To address the problem of data scarcity, synthetic data can be used to replace or substitute missing data [SQLG15, HPH$^+$19, RVRK16]. Synthetic data has many advantages that overcome the problems associated with real datasets. One of the main advantages is the ability to generate large amounts of data with automatic labelling. In this sense, simulators enable experimental conditions to be controlled, thus avoiding most annotation or labelling errors. For ML research -such as that conducted as part of the Frugal And Robust AI for Defence Advanced Intelligence (FaRADAI) project- where the choice of training data is crucial for algorithm calibration, synthetic data represents an interesting opportunity to generate specific use cases for algorithm training. Synthetic data simulators enable the creation of large datasets to train ML algorithms at a lower cost and in less time than real data. Their advantages stem from their ability to repeat simulations and control simulator parameters. Thanks to these capabilities, these simulators are versatile and can be used for many different tasks, such as generating the training datasets for regression learning . This advantage is crucial in the defence sector for processing enemy RF signatures, as this technology offers the possibility of simulating this type of data.

However, generating synthetic data for the purpose of training a DL algorithm requires a non-naive construction procedure. Without caution, a synthetic dataset can introduce biases or unrealistic samples, leading to inference problems [ERV$^+$23]. One of the main drawbacks of using synthetic data compared to real data is the discrepancy between synthetic and real data, at the individual or global level. The difficulty of the problem lies in defining the level of realism of synthetic data. Many researches been conducted to define rules for quantifying the realism of synthetic data [Tzi22, SZY23]. Measuring the quality of datasets is essential for synthetic dataset generators [JPN$^+$20]. Most researches focus on comparing real and synthetic datasets. However, in our case, we are trying to find a solution to determine the quality of a synthetic dataset without any real data as a reference. To do this, we gather different rules, illustrated by different types of metrics, which serve as indicators. Six categories of arithmetic measures give immediate results on the dataset. They can be considered as a greedy algorithm that roughly analyses the dataset in order to eliminate those which are of insufficient quality in terms of image resolution, contrast, luminance, sharpness... Then, the other categories, based on analysis by neural models, attempt to evaluate the diversity of the images and the quality of the different situations illustrated in the images for a detection task.

In this article, after an introduction section presenting a brief state of art about the use of synthetic data in the ML filed, section 2 presents the simulation toolkit SE-WB from which all synthetic data in this study was generated. We then present the prototype with its different categories of measurements, classified according to whether they are arithmetic or neural analyses. The last section examines the results of the various experiments conducted as part of this study and concludes with a summary.

2 SE-WORKBENCH-EO

SE-WORKBENCH, also called CHORALE in France, is a toolkit developed by the French company Oktal-SE to address the problem of generating sensor outputs in different domains such as visible, IR[1] domains and for RF systems such as RAdio Detection And Ranging (RADAR) systems. With its RF and EO versions, this set of tools aims to model the $3D$ environment using a virtual and synthetic approach to render the $3D$ scene for a given sensor. Its main applications are focused on parametric studies. Simulation scenarios are constructed from the interweaving of several blocks, mainly object and terrain modelling. The materials created in the simulator have physical properties based on real measurements. Thermal modelling is also applied according to the different elements of the scene, and atmospheric modelling (MODerate resolution atmospheric TRANsmission (MODTRAN)) is also implemented.

3 CLEANSE

3.1 Principle

Checking Learning Entries and Anomalies for Network Stability and Efficiency (CLEANSE) is a prototype designed to validate synthetic data generated by the SE-WB toolkit. Those data are intended specifically for training DL algorithms. CLEANSE uses a series of tests to measure and quantify the conceptual consistency, diversity and quality of the data. In our study, we limited our work to synthetic images intended for detection models. The metrics assign individual scores, which are then combined to calculate an overall score. This overall score indicates the suitability of the SE-WB data for training DL models. CLEANSE aims to improve synthetic data for the training phase of DL models. This paper presents the first version of CLEANSE. The experiments presented below establish the basis for qualitative and quantitative evaluations of synthetic data.

[1]from band I to band III

To achieve this,CLEANSE will rely on two types of metrics: arithmetic and neural.

3.2 Arithmetic metrics

The first set of metrics presented are arithmetic metrics. These metrics enable a conceptual analysis of the dataset to be performed. They provide an immediate overview of the dataset's average quality. Although based on relatively 'less complex' metrics than neural metrics, they are a first step in filtering out irrelevant datasets before analysing them in greater depth. They can also be used to detect anomalies in the dataset, such as blurred images or duplicates, but this is not their primary function.

Arithmetic metrics cover several categories, of which there are six in CLEANSE v0.

3.2.1 Category 1: Structural properties and image metadata

This category focuses on image characteristics such as the variety of colours displayed and the image format. The information is provided by the following metrics:

- Number of distinct colours N_c, which provides information on the visual richness and compression effects of an image,
- Colourfulness is a metric that measures the vividness of colours as perceived by the human eye. In this case, the metric is based on subjective criteria and attempts to quantify the chromatic overload and content of the image,
- Perceptual Hash (pHash): is therefore a coding of the overall visual structure. Images with identical pHash values are highly likely to be visual duplicates.

3.2.2 Category 2: Luminance and intensity dynamics

This category of measures attempts to quantify the overall brightness, contrast, and distribution of hues in an image.

- Average intensity μ: the average intensity of an image I, also noted μ_I, is the average value of its grey levels. The image I is first converted into its monochrome version (grey levels) before the average intensity is computed. Given that $I = (I_{x,y}) \in \mathcal{M}_{n,p}(\mathbb{K})$, where $(n, p) \in \mathbb{N}^2$ are the image dimensions, we have:

$$\mu_I = \frac{1}{np} \sum_{x=1}^{n} \sum_{y=1}^{p} I_{x,y}. \tag{1}$$

  The closer μ_I is to 0, the darker the image I is. Conversely, a value of μ_I close to 255 qualifies an image I as very clear. Thus, μ_I provides information about the exposure in image I,

- Standard deviation of intensities σ: based on an image I converted to greyscale, this is a measurement that quantifies the contrast level of an image, computed as follows:

$$\sigma_I = \sqrt{\frac{1}{np} \sum_{x=1}^{n} \sum_{y=1}^{p} (I_{x,y} - \mu_I)^2}. \tag{2}$$

3.2.3 Category 3: Edge sharpness

In this category, the sharpness of contours within an image is measured to determine the level of blur:

- Laplacian variance is used to estimate the presence of sharp contours. Also known as sharpness, the sharpness Sh of an image I is computed as $Sh_I = Var(\nabla^2 I)$. A low value of Sh_I implies the presence of few sharp contours, as in the case of a blurred image,

- The Tenengrad metric measures the sharpness of an image I based on the strength of its contours. With G_x^I the gradient of I on the x-axis and G_y^I the gradient on the y-axis, the Tenengrad metric T is defined as follows:

$$T = \sum_{1 \leq i \leq n, 1 \leq j \leq p} \underbrace{\left(G_x^I(i,j)\right)^2 + \left(G_y^I(i,j)\right)^2}_{\text{the square of the gradient energy}} \tag{3}$$

- Sum of Modified Laplacian (SML): measuring sharpness in noisy conditions. The image I is convolved with a modified Laplacian and only the significant intensities are retained for addition.

3.2.4 Category 4: Local texture

In this category, the introduced metrics scale the complexity of existing patterns within an image. The Gray Level Co-occurrence Matrix (GLCM) is built by classifying image pixels and comparing their grey level to that of their neighbours [Gad04]. After normalisation, the resulting matrix, $\hat{GM}$, becomes a probability distribution. Information about the smoothing, regularity, transitions and patterns of the image can be extracted from the matrix.

- Image contrast is measured using the local variance, V_l, which measures the deviation of grey levels from the computed average.

$$V_l = \sum_{i=1}^{N_{GM}} \sum_{j=1}^{P_{GM}} (i - \nu)^2 \hat{GM}_{i,j} \tag{4}$$

  where

  - $\nu_x = \sum_{i=1}^{N_{GM}} \sum_{j=1}^{P_{GM}} i.\hat{GM}_{i,j}$ the marginal average in the x-axis,
  - N_{GM} and P_{GM} are the dimensions of the GM matrix.

  A weak V_l is synonymous with weak hue transitions, which are typical of uniform image textures,

- Angular Second Moment (ASM) of an image: to note whether the matrix is concentrated around certain values. The ASM, also known as energy, is measured as follows:

$$ASM = \sum_{i=1}^{N_{GM}} \sum_{j=1}^{P_{GM}} (\hat{GM}_{i,j})^2 \tag{5}$$

  The energy of GM will be maximum for a regular/periodic image,

- Homogeneity, which quantifies the proximity between neighbours, is defined as follows:

$$h = \sum_{i=1}^{N_{GM}} \sum_{j=1}^{P_{GM}} \frac{\hat{GM}_{i,j}}{1+(i-j)^2}$$

. High homogeneity will be observed when neighbours are similar, which is typical of a smooth image,

- Correlation between neighbouring grey levels. A correlation close to 1 would signify progressive and structured patterns in the image, while a correlation close to 0 would indicate random textures. The correlation can have a negative value, which would indicate inverted patterns. The correlation C is defined as follows:

$$C = \sum_{i=1}^{N_{GM}} \sum_{j=1}^{P_{GM}} \frac{(i-\nu_x)(j-\nu_y)\hat{GM}_{i,j}}{\sigma_x \sigma_y} \quad (6)$$

where

  - $\nu_y = \sum_{i=1}^{N_{GM}} \sum_{j=1}^{P_{GM}} j.\hat{GM}_{i,j}$ the marginal average in the y-axis,
  - $\sigma_x = \sqrt{\sum_{i=1}^{N_{GM}} \sum_{j=1}^{P_{GM}} (i-\nu_x)^2 \hat{GM}_{i,j}}$,
  - $\sigma_y = \sqrt{\sum_{i=1}^{N_{GM}} \sum_{j=1}^{P_{GM}} (j-\nu_y)^2 \hat{GM}_{i,j}}$.

3.2.5 Category 5: Frequency analysis

This category focuses on the frequency aspect of images, particularly their distribution across the spectrum. For example, a sharp image preserves energy at high frequencies, whereas a blurred image concentrates energy at low frequencies:

- The Spectral Energy Ratio (SER) is computed applying a $2D$ Fourier Transform of a monochrome image I:

$$S(u,v) = \int_{x=-\infty}^{x=\infty} \int_{y=-\infty}^{y=\infty} I(x,y).e^{-2j\pi(ux+vy)} dxdy$$

with $j^2 = -1$. Then, with the spectrum S we compute

$$SER_{HF} = \frac{\sum_{(u,v)\in HF} S(u,v)}{\sum_{u,v} S(u,v)} \quad (7)$$

where $HF = \left\{(u,v)\ \sqrt{u^2+v^2} \geq R \text{ with } R > 0\right\}$ the set of high frequencies. To avoid overloading the notation, we will refer to SER_{HF} as SER,

- Spectral Entropy (SE): With $\hat{S}$ the normalized spectrum, the SE is computed as:

$$H_{Spec} = -\sum_{u=1}^{N_{\hat{S}}} \sum_{v=1}^{P_{\hat{S}}} \hat{S}(u,v).\log_2\left(\hat{S}(u,v)\right) \quad (8)$$

with $N_{\hat{S}}$ and $P_{\hat{S}}$ the dimension of the normalized spectrum $\hat{S}$. A regular image will have a weak SE, while a noisy image or one with complex patterns will have a high SE.

3.2.6 Category 6: Global content statistics

The analysis in this final category of arithmetic metrics is based on the observation of the distribution of grey levels.

- The Spatial Entropy measures the richness of the visual information and is defined as:

$$H_{Spat} = -\sum_{i=0}^{255} h_i.\log_2(h_i) \quad (9)$$

with $(h_i)_{0\leq i\leq 255}$ the histograms of the monochrome image I. A high spatial entropy value indicates a balanced distribution of image hues, which is typical of an image with a complex pattern. In contrast, a low spatial entropy value indicates a uniform or compressed image,

- The Kurtosis metric, defined as:

$$Kurtosis = \frac{1}{N} \sum_i \left(\frac{x_i - \mu_I}{\sigma_I}\right)^4 \quad (10)$$

with N the number of pixels. A Kurtosis value greater than 3 indicates pronounced peaks, which are linked to noise, compression or detail in an image, while a Kurtosis value lower than 3 corresponds to a flattened distribution of grey levels, as seen in a blurred image.

3.3 Neuronal based metrics

The goal of the CLEANSE v0 neural metrics is to provide a deeper analysis of datasets validated by arithmetic metrics. In this set of metrics, data is fed into a neural model to compute metrics from the model outputs or to extract information from the different feature maps of the hidden layers. These metrics can be divided into two groups: models that evaluate the quality of the dataset, and models that evaluate its diversity. We expect these metrics to measure the compatibility of synthetic data with the requirements for DL algorithm training. These metrics will also demonstrate whether the data generated by the SE-WB simulator is sufficiently diverse and qualified for DL applications.

The first category focuses on dataset quality, while the second category focuses on dataset diversity and the local and global variability of the data. In this context, variability poverty poses a threat to DL training due to the possibility of overfitting [CT10, SK19]. Conversely, it is not advisable to have too much variability in the dataset if the neural model is to converge in a finite amount of time [YAH23, KMR20, LSAS23].

3.3.1 Quality

In this first category, we introduce neuronal metrics and a better procedure based on neural models to estimate the quality of a dataset without references, such as real-world data (unless for Natural Image Quality Evaluator (NIQE)):

- Constrastive Language-Image Pretraining Image Quality Assessment (CLIP-IQA) is a no-reference, zero-shot image quality metric introduced by Wang et al. in [WCL23]. It uses the pre-trained Constrastive Language-Image Pretraining (CLIP) model to project images and text into a shared semantic space, and encodes an input image using the CLIP visual encoder. It then compares the encoded image to carefully designed text prompts describing different levels of quality (e.g. 'a sharp picture', 'a blurry picture') via cosine similarity. The relative similarities between the image embedding and these quality-related text embeddings are converted into a quality score. Unlike supervised Image Quality Assessment (IQA) models, CLIP-IQA does not rely on IQA ground truth during inference, and its zero-shot variant does not require retraining on IQA datasets,
- Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [MMB12] works with Scene Natural Statistics (SNS). These are computed using Mean Subsracted Contrast Normalization (MSCN) on the image, which provides Gaussian distributions whose characteristics are modified according to distortions found in the image. It also performs local normalization on the image, eliminating local lighting variations and highlighting the local structures of the image. Distortions can provide information about neighbouring pixel relationships and local dependencies within the image. MSCN cross products in the four directions (horizontal, vertical, and along the diagonals) enable other distributions sensitive to the image's structure and texture to be extracted. Once this initial step has been completed, the MSCN and its cross products are computed again on the subsampled version of the image. The characteristics of the different distributions of the full and subsampled images are then gathered into a single vector of characteristics. This vector is then applied to a Support Vector Machine (SVM) pretrained on an image database labelled following human evaluation of quality. The score depicted by the SVM is the level of image quality estimated by BRISQUE. The different steps of the BRISQUE quality evaluation process are illustrated in Figure 1,
- Like BRISQUE, NIQE [MSB12] bases its image analysis on SNS. However, in the case of NIQE, there is no pre-training with human-annotated data. The idea behind NIQE is that natural images have specific statistical uniformities that are distorted when the image is degraded. NIQE estimates visual quality by computing the distance between the image to be analysed and a set of referenced images considered to be of high

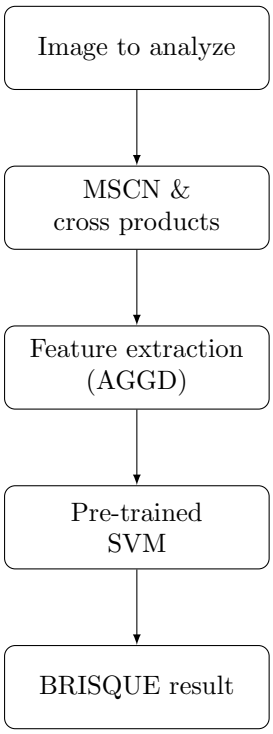

Figure 1: Image quality analysis with BRISQUE neural metric.

quality. As with BRISQUE, MSCN and its cross products are computed, but different features are extracted. Then, on the vector of characteristics, a Mahalanobis distance is computed instead of a SVM application, which enables the NIQE score to be deduced:

$$NIQE(I) = \sqrt{(f_I - \mu_{nat})\Sigma^{-1}(f_I - \mu_{nat})} \quad (11)$$

with

- I the image whom quality to evaluate,
- f_I the vector of characteristics of I,
- The covariance matrix, denoted by Σ, describes the variance and correlations between the characteristics,
- μ_{nat} averaged vector of characteristics for the reference natural image dataset.

The Mahalanobis distance is a measure of how far an image deviates from the expected behaviour of natural images used as references.

3.3.2 Diversity

The following metrics aim to evaluate the balance between internal coherence and the richness of the dataset:

- CLIP Embedding Distance: in this metric, images are projected into a vectorial space in order to measure the dataset's overall diversity. Similar visual or semantic concepts will have similar positions in the projected space. This is achieved by computing an average distance between each pair of embeddings:

$$d_{avg} = \frac{1}{Card(Emb)} \sum_{(A,B)\in Emb} \underbrace{\frac{AB}{||A||.||B||}}_{\cos(A,B)} \quad (12)$$

with

  - (A, B) a pair of embedding vectors,
  - Emb the embedding vector set,
  - $Card(X)$ the dimension for a field X,
  - $. \mapsto ||.||$ a norm function defined in the vectorial space.

  CLIP Embedding distance identifies high-level differences, such as the presence of different object categories or visual patterns, in the dataset.

- Learned Perceptual Image Patch Similarity (LPIPS) introduced in [ZIE+18] is a metric that evaluates the similarity between images based on how similar the feature maps of a Convolutional Neural Network (CNN) are. To compare two images I_1 and I_2 six steps must be followed:

  1. Forward of I_1 and I_2 in the CNN,
  2. Feature maps extraction for each CNN convolutional layer. $\forall l \in L$, L being the set of layers with feature maps:

  $$F_l(I_1) \in \mathbb{R}^{H_l \times W_l \times C_l}$$

  is the activation at the layer l with

- H_l the height of the feature map at the layer l,
- W_l the width of the feature map at the layer l,
- C_l the number of channels at the layer l.

3. Feature map which normalisation reduces the influence of magnitude variation to focus on the relative structure of layer activations:

$$\hat{F_{l,c}}(I_1) = \frac{F_{l,c}(I_1)}{||F_{l,c}(I_1)||_2}$$

with

- $. \mapsto ||.||_2$ a spatial norm per channel,
- $c \in C_l$ a channel of the feature map in layer l.

4. Distance computation between pairs of feature maps for each channel is performed. For a channel $c \in C_l$ at layer $l \in L$:

$$d_{l,c}(I_1, I_2) = ||\hat{F}_l(I_1) - \hat{F}_l(I_2)||_2^2$$

5. Weighting of each channel with the CNN weights (to weight their contribution in the visual perception):

$$LPIPS_l = \sum_{c=1}^{C_l} \omega_{l,c}.d_{l,c}$$

6. Obtained LPIPS score:

$$LPIPS(I_1, I_2) = \sum_{l \in L} LPIPS_l \qquad (13)$$

The LPIPS value, contained within the interval $[0, +\infty[$, provides an indication of the global homogeneity of a set of images. A LPIPS value lower than 0.05 indicates that the two tested images are nearly indistinguishable, whereas a LPIPS value higher than 0.5 indicates that the two images have sharp and observable differences,

- The Vendi Score, introduced by Friedman et al. in [FD22] counts the number of distinct spectral information states in an image. To achieve this, a similarity matrix $M \in S_n(\mathbb{K})$ with $n \in \mathbb{N}^*$ elements is constructed, n being the number of data. Then the distribution of similarities, $\hat{M} = M/n$, is defined. The diversity of the set of n samples is defined by computing the eigenvalues of M, and the Vendi Score VS is defined based on the Shannon entropy formula:

$$VS = \exp\big(-\sum_{i=1}^{n} \lambda_i \log(\lambda_i)\big) \qquad (14)$$

The Vendi Score examines the global structure of the dataset and calculates the spectral diffusion of the eigenvalues. If $rg(\hat{M}) = 1$ and $VS = 1$, then all the dataset samples are identical, whereas $\hat{M} \in D_n(\mathbb{K})$ and $VS = n$ indicate that all the data are orthogonal,

- The Random Network Distillation (RND) metric involves measuring the prediction error between two Neural Network (NN). The first NN, f_1, is randomly initialised and frozen. Then a second neural network, f_2, is trained to imitate the output of f_1. Therefore, the prediction error for an image I is

$$err_{pred}(I) = ||f_1(I) - f_2(I)||^2$$

The prediction error measures the novelty of a sample. For data that has never been seen before, the error will be high; therefore, the cumulative error on the whole dataset indicates the diversity of the dataset. The dataset is divided into two subsets for training and validation, and the RND is obtained as follows:

$$RND = \frac{loss_{val} - loss_{train}}{loss_{train}} \qquad (15)$$

with $loss_{val}$ and $loss_{train}$ respectively the loss on the validation and training set,

- As with previous neural metrics, kNN Latent Distance metric projects an image I into a latent

space LS to measure distances between embeddings. It measures novelty by checking whether data in the latent space look the same and measures diversity by checking whether data are concentrated around certain areas. The embeddings are computed by forwarding the data through a NN. To define a notion of distance in the latent space LS of dimension $d \in \mathbb{N}^*$, a R-sphere with radius $R > 0$ is built around a data point $z_g \in LS$ to search for the k-nearest neighbours ($k > 0$): $k_{NN}(z_g, R)$. Then, the average distance is defined as:

$$d_g^R = \frac{1}{k} \sum_{z_j \in k_{NN}(z_g,R)} ||z_g - z_j||$$

and the average error on the whole dataset D is computed as follow:

$$k_{NN-LD}(D, R) = \frac{1}{Card(D)} \sum_{g \in D} d_g^R \qquad (16)$$

A weak k_{NN-LD} value suggests that the dataset is lacking in novelty, whereas a high value indicates that it is diverse,

- Self-FID (Self-FID) is a variant of Frechet Inception Distance (FID) without reference data, which aims to measure the internal diversity of a dataset. It extracts the feature map for a dataset from a NN. The group of images is split to enable comparison with each other. For each group, the average and covariance of the activation vectors are computed and the metric is calculated:

$$Self-FID = ||\mu_1 - \mu_2||_2^2 + Tr\big(\Sigma_1 + \Sigma_2 - 2(\Sigma 1 \Sigma_2)^{\frac{1}{2}}\big) \qquad (17)$$

The dataset is considered rich if the neural model generates a lot of modes.

4 Experiments

Several synthetic datasets produced from SE-WB were used for training and inference processes in a You Only Look Once (YOLO) v7 algorithm [WBL23]. The aim of the experiments is to confirm the score obtained by the metrics with the training and generalisation performance of the YOLO algorithm for plane detection. Several datasets of 3,000 samples each were built for the experiments. Of these, 500 were generated without targets to detect:

Table 1: Weighting of the base datasets used to construct datasets C to G.

Dataset built	Dataset A	Dataset B	Negative Dataset
Dataset C	1/3	2/3	$\emptyset$
Dataset D	1/2	1/2	$\emptyset$
Dataset E	2/3	1/3	$\emptyset$
Dataset F	2/3	$\emptyset$	1/3
Dataset G	1/3	$\emptyset$	2/3

- Dataset A was built to be diverse and of high quality. Several types of target were used in this dataset, including civilian and military planes. The images were generated in several $3D$ synthetic airport environments, including scenes based on ortho-images of the Blagnac, Chambéry and Kansai airports in France and Japan, respectively. Various weather conditions were simulated, including foggy, sunny and cloudy scenarios, with sensor images captured at different times of day (morning, midday, afternoon and evening),

- Dataset B was deliberately created to be poor in terms of diversity, with only one target (an A320 civil aircraft) at Chambéry airport at midday in sunny conditions.

Several merges were performed between datasets A and B to create datasets with an increasing level of diversity and content quality. These datasets and their content are summarised in Table 1, with the 'Negative Dataset' being a dataset generated without planes, but with other objects instead.

4.1 YOLOv7 metrics

To observe the performances of the YOLO algorithm used, 6 key metrics will be used:

- Box loss is the error measurement made on box localisation. Ideally, this should be as low as possible,
- Precision is the percentage of correct detections among all detections. It must be above 80%,
- The recall is the percentage of detections among all the images containing the objects to be detected. It must be above 80%,
- The $F1$ score is the balance between precision and recall at a given confidence level. Ideally, the peak is between 0.5 and 0.6,
- mAP@0.5: precision average for a Intersection over Union (IoU) over 50%. Ideally, it should be above 75%,
- mAP@0.5:0.95: strict average for several IoU from 0.5 to 0.95. Ideally, it should be above 75%.

4.2 Results

Tables 2 and 3 refer to the arithmetic metrics applied on the different datasets while tables 4 and 5 contain the neural metrics for diversity and quality scores obtained by the different datasets. For tables 2, 3 and 5, as the metric were computed for each image of the dataset, we will only display the average with the standard deviation on the whole dataset. Also, table 6 gives the YOLO performance indicator.

Different behaviour can be observed in Tables 2 and 3. While metrics such as spatial entropy (H_{Spat}) or correlation seem to be almost identical regardless of the dataset, metrics such as the number of distinct colours (N_c), sharpness and the Tenengrad metric differ. From datasets A to B (Table 2), colourfulness increases, meaning an enhancement of colour brightness with the loss of dataset diversity. However, for datasets F and G, the colourfulness is equivalent to that of a diverse dataset such as dataset A or E. Even though the figures are of the same order of magnitude, the average intensity of greyscales in the image (μ_I) increases when dataset diversity decreases. Energy and homogeneity decrease with diversity, indicating more uniform and subtle transitions between pixels in more diverse datasets (A or E) than in less diverse datasets (C or B). According to the SER score, poorer datasets are found to be more distinct. Kurtosis increases with poorer datasets. As all Kurtosis values are below 3, indicating weak Kurtosis, all datasets are considered to have a soft grey distribution in the images (i.e. a blurring effect).

Another notable observation regarding the two arithmetic tables is that the scores for datasets F and G are often similar to those for datasets A and E, which were designed to be the most diverse in the experiments of this study.

Table 4 shows that, firstly, the CLIP embedding distance score decreases when the dataset is poorer in diversity. However, it can be seen that datasets with 'negative samples' increase the CLIP score, reaching the level of datasets D and E. The same can be observed for LPIPS and Vendi scores. For Self-FID, however, Dataset A obtained a lower score, close to that of Dataset B. These scores indicate that homogeneous diversity has been detected in both the most diverse Dataset (A) and the least diverse Dataset (B). For k_{NN-LD}, Dataset A and Dataset F obtained the best scores, while Dataset B, the dataset with the worst diversity, obtained the weakest score. As with CLIP, the RND scores are close to each other. According to this metric, Dataset B contains the most novel content of all the datasets analysed.

When looking at the quality measurements provided by the neural metrics in Table 5, similar values of CLIP-IQA and BRISQUE can be observed for Datasets A and F and G. Excluding these two datasets, CLIP-IQA increases when the dataset quality is poorer, even though it is still far from perfect (CLIP-IQA = 1). According to BRISQUE, the poorer the dataset, the less degraded the image is considered on average. NIQE also indicates a better quality score for datasets with poorer diversity, such as B, than for datasets built for greater diversity, such as E or A.

Table 6 gives the figures of the YOLO v7 algorithm performances when trained with the different data. While the error made on the box location is higher for more diverse datasets, precision and recall drops

[3]Negative $pHash$ means that no duplicates found in the tested dataset

Table 2: Arithmetic metrics for dataset evaluation, starting from the more diverse to the less one.

Category	Metric	Dataset A	Dataset E	Dataset D	Dataset C	Dataset B
I	Colorfullness	12.77 ± 6.73	15.11 ± 6.50	16.23 ± 6.06	17.42 ± 5.30	19.69 ± 2.34
	N_c (10^3)	5.5 ± 3.9	8.6 ± 5.7	10.2 ± 5.8	11.7 ± 5.6	14.6 ± 3.2
	pHash	Negative[2]				
II	μ_I (10^2)	1.15 ± 0.39	1.22 ± 0.34	1.25 ± 0.30	1.29 ± 0.26	1.36 ± 0.12
	σ_I	54.14	23.93	23.93	23.93	18.67
III	Sharpness (10^2)	0.82 ± 0.82	1.39 ± 1.15	1.69 ± 1.18	1.99 ± 1.14	2.57 ± 0.76
	Tenengrad (10^3)	0.90 ± 0.9	1.7 ± 1.6	2.1 ± 1.5	2.4 ± 1.5	3.22 ± 1.07
	SML (10^6)	1.09 ± 0.63	1.67 ± 0.11	1.96 ± 1.16	2.26 ± 1.10	2.85 ± = 0.75
IV	Contrast	0.47 ± 0.51	0.69 ± 0.60	0.82 ± 0.64	0.94 ± 0.63	1.18 ± 0.55
	Energy	0.29 ± 0.08	0.26 ± 0.08	0.25 ± 0.07	0.24 ± 0.07	0.21 ± 0.03
	Homogeneity	0.95 ± 0.03	0.92 ± 0.06	0.90 ± 0.06	0.89 ± 0.06	0.86 ± 0.04
	Correlation	1.00 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01
V	SER (10^{-3})	0.8 ± 0.6	0.9 ± 0.6	1.0 ± 0.5	1.0 ± 0.5	1.2 ± 0.4
	H_{Spec}	0.88 ± 0.45	0.80 ± 0.40	0.76 ± 0.36	0.72 ± 0.33	0.65 ± 0.21
VI	Kurtosis	-1.02 ± 2.03	-0.35 ± 2.61	-0.03 ± 2.63	0.30 ± 2.61	0.96 ± 2.44
	H_{Spat}	5.54 ± 0.18	5.54 ± 0.14	5.54 ± 0.18	5.54 ± 0.18	5.54 ± 0.12

Table 3: Arithmetic metrics for dataset (F & G) evaluation.

Category	Metric	Dataset F	Dataset G
I	Colorfullness	13.38 ± 6.60	13.37 ± 6.83
	N_c (10^3)	5.6 ± 3.9	5.3 ± 4.0
	pHash	Negative[3]	
II	μ_I (10^2)	1.20 ± 0.39	1.18 ± 0.38
	σ_I	36.63	15.05
III	Sharpness (10^2)	0.90 ± 0.89	0.94 ± 1.04
	Tenengrad (10^3)	1.00 ± 1.02	1.06 ± 1.24
	SML (10^6)	1.12 ± 0.66	1.12 ± 0.67
IV	Contrast	0.48 ± 0.54	0.47 ± 0.57
	Energy	0.29 ± 0.08	0.30 ± 0.08
	Homogeneity	0.95 ± 0.04	0.95 ± 0.04
	Correlation	1.00 ± 0.01	1.00 ± 0.01
V	SER (10^{-3})	0.77 ± 0.56	0.77 ± 0.56
	H_{Spec}	0.86 ± 0.43	0.87 ± 0.43
VI	Kurtosis	-1.01 ± 2.44	-1.07 ± 2.24
	H_{Spat}	5.54 ± 0.12	5.54 ± 1.19

Table 4: Neural metrics to evaluate dataset diversity.

Neural metrics	Dataset A	Dataset E	Dataset D	Dataset C	Dataset B	Dataset F	Dataset G
CLIP (10^{-1})	1.6	1.6	1.4	1.3	0.9	1.5	1.4
LPIPS (10^{-1})	5.11	5.05	4.96	4.80	4.24	4.96	4.90
Vendi Score	2.74	2.64	2.49	2.31	1.79	2.65	2.50
Self-FID (10^{-3})	3.1	4.7	4.4	4.2	3.7	6.2	5.0
k_{NN-LD} (10^{-2})	3.2	3.0	2.7	2.3	1.4	3.4	2.2
RND (10^{-2})	1.181	1.113	1.171	1.082	1.214	1.145	1.185

Table 5: Neural metrics to evaluate dataset quality.

Neural metrics	Dataset A	Dataset E	Dataset D	Dataset C	Dataset B	Dataset F	Dataset G
CLIP-IQA (10^{-1})	2.63 ± 1.38	3.29 ± 1.56	3.61 ± 1.54	3.94 ± 1.45	4.60 ± 0.97	2.54 ± 1.36	2.31 ± 1.37
BRISQUE (10^{1})	5.77 ± 1.78	4.84 ± 1.99	4.36 ± 1.93	3.91 ± 1.77	2.96 ± 0.66	5.79 ± 1.86	5.86 ± 1.89
NIQE (10^{1})	1.09 ± 0.23	1.04 ± 0.22	1.01 ± 0.21	1.00 ± 0.20	0.95 ± 0.16	1.10 ± 0.23	1.12 ± 0.23

significatively when the dataset is poorer. Interesting point to note, the precision obtained by the model on the dataset F is equal to the precision obtained on the more diverse dataset (A). Similar observations can be remarked on the different mAP@.

4.3 Analysis

Table 4 shows that the different scores reveal some interesting observations. As expected, CLIP and LPIPS increase when the dataset contains more diversity. As we built Dataset B, Dataset C, Dataset D and Dataset E from poorer to more diverse datasets, the two metrics highlight the increasing diversity of the datasets. However, it should be noted that Datasets F and G obtain scores close to Datasets E and D, respectively, meaning that adding 'negative samples' of objects that do not need to be detected improves the diversity of the dataset. This similarity in the scores obtained by Datasets A and E, and Datasets F and G, is also observable in tables 2, 3 and 5. Additionally, some metrics seem unable to distinguish between the analysed datasets, such as the correlation and spatial entropy metrics. According to spatial entropy H_{Spat}, for example, the distribution of objects in the images of the different datasets is similar. While the arithmetic metrics give more credit to the quality of the less diverse datasets, in Table 6 we observe that diversity has a significant impact on the precision of the model's detection, with a 26% drop in precision between Datasets A and B. Also, we observe that adding too many 'negative samples' to Dataset G poisons the dataset and compromises the learning of the YOLO algorithm.

Generalizing what can be observed, the quality metrics indicate that in a synthetic dataset generated by the SE-WB, the quality of the images is better perceived when working with a 'less' diverse dataset. Metrics like Sharpness or Tenengrad indicate that Dataset B images look clearer and contain more energy in their gradients than images from Dataset A. However the standard deviation of intensities σ_I returns decreasing values from Dataset A $\rightarrow$ B, telling that global contrast is higher in more diverse dataset. Something explainable by the different hours of image generation from the morning to the evening contrary to Dataset B only generated for midday conditions. As Dataset F as a lot of common images with Dataset A, σ_I for F is higher than for every other dataset except the Dataset A. As the dataset were built with a focus on diversity, it is not surprising to obtain quality scores favoring the poorer datasets. However, the results can also lead to other explanations such as the definition of quality in the metrics. For the neural metrics notably, many of them rely on neural models for the embedding coding of images for instance. In

Table 6: YOLO metrics to evaluate its performances when trained with a dataset

YOLO metrics	Dataset A	Dataset E	Dataset D	Dataset C	Dataset B	Dataset F	Dataset G
Box loss (10^{-2})	2.0	1.9	2.0	2.1	1.5	1.6	5.0
Precision (%)	88	85	84	81	62	88	7
Recall (%)	82	80	64	61	31	70	22
$F1$ score (10^{-1})	8.5	8.2	7.3	7.0	4.1	7.8	1.1
mAP@0.5 (10^{-1})	8.2	8.15	8.0	6.3	3.1	8.0	0.4
mAP@0.5:0.95 (10^{-1})	6.1	6.1	5.1	4.1	1.7	5.3	0.08

this way, the generated images, even with diversity, could have generated images further away from the characteristics of images learnt by the different models. The high standard deviation can also explain conflicting scores such as μ_I indicating more intensities in the pixel of the images in the poorer datasets while σ_I show the contrary.

4.4 Identification of the metric limits

The neuronal metrics results must be analysed carefully. Unlike the arithmetic metrics, the neuronal metrics results depend on various factors, such as the neural model used for image embedding in the k-nearest neighbour latent distance or affected by the human annotations in BRISQUE. However, they are necessary to provide information on the quality and diversity of image datasets. That is why CLEANSE is organised with metrics from different domains, to extract as much information as possible and reduce the bias and evaluation errors made by the different metrics. For example, the three neural quality metrics were applied to one real image and one from SE-WB. These two images can be seen in Figure 2, while the results of the different metrics are displayed in Table 7. For the CLIP-IQA and BRISQUE metrics, the real image obtained a better score than the synthetic image. For CLIP-IQA, synthetic data obtained a weaker score than real data. However, this does not necessarily mean that the image is of a lower quality. The score also suggests that the synthetic data is far from the conventional quality concepts learnt by the CLIP-IQA model. The same applies to BRISQUE: the synthetic data may not correspond to the characteristics expected by the model and does not necessarily imply that the image is degraded. For the NIQE result that goes against the other two, the score can be explained by better internal coherence within the synthetic image compared with the real image, even though the latter is considered of better quality by the other two metrics.

Neural metric	Score	
	Image A	Image B
CLIP-IQA (10^{-1})	7.2	4.5
BRISQUE	38.0	52.9
NIQE	11.9	10.4

Table 7: Score table for the neural quality metrics on the A & B images.

5 Conclusion

CLEANSE v0 is a pioneering prototype designed to assess the quality and diversity of synthetic datasets used for training DL models. In cases of data scarcity, evaluating data without reference is a complex task. To address this issue, the first prototype brought together a variety of different indicators. These were separated into two groups. The first group contains only arithmetic metrics (section 3.2) for rapid, resource-intensive statistical analysis. The second group comprises all metrics based on neural metrics section 3.3) for deeper analysis about the internal structure, texture and coherence of the dataset. Experiments were conducted to observe and compare the scores obtained by the different metrics with the performance of a YOLO v7 algorithm. This procedure established a link between statistical visual anal-

Image A Image B

Figure 2: Image from a real EO sensor in an airport environment and simulated visible EO sensor output in a $3D$ airport synthetic environment.

ysis and DL performance. The experiments yielded interesting results. As expected, there were numerous conflicts between the metrics, which is unsurprising given the large number of measurements used (25, not counting the YOLO metrics). Generally, a high standard deviation was observed, rendering this measure and the tested datasets unreliable. However, some measures, such as the k-NN Latent Distance and the Vendi Score, appear to reflect the diversity present in the data.

The first steps of CLEANSE are encouraging, but need to be improved. As this is a first prototype, we expected many metrics to be untrustworthy, which appears to be the case for many arithmetic metrics notably. However, before that, CLEANSE v0 should be tested on other datasets with degraded or enhanced quality, as well as on other simulations, such as in other synthetic environments or for different objects. As mentioned in section 4.4, many neural metrics can be influenced by adjustable factors. These were fixed for the initial experiments and would require re-evaluation in future tests. Additionally, the quality of the synthetic data generated by SE-WB may not be sufficient for neural metrics that rely on natural or high-quality images for pre-training (NIQE metric for instance). Finally, the YOLO performances show that a diverse synthetic dataset generated by the SE-WB toolkit can successfully train a DL algorithm, as can be seen from the precision, recall, $F1$ score and mAP@ metrics in Table 6.

Acknowlegments

This study was supported by the European Defence Fund project FaRADAI. Main research thematics concern the adaptations on new environments and objects/Enemy identification and recognition, the Robustness and explainability of Automated Target Detection (ATD) and Automated Target Recognition (ATR).

References

[CT10] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

[ERV+23] Peter Eigenschink, Thomas Reutterer, Stefan Vamosi, Ralf Vamosi, Chang Sun, and Klaudius Kalcher. Deep generative models for synthetic data: A survey. *IEEE Access*, 11:47304–47320, 2023.

[FD22] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

[Gad04] Dhanashree Gadkari. Image quality analysis using glcm. 2004.

[HPH+19] Stefan Hinterstoisser, Olivier Pauly, Hauke Heibel, Marek Martina, and Martin Bokeloh. An annotation saved is an annotation earned: Using fully synthetic training for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[JPN+20] Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3561–3562, 2020.

[KMR20] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International conference on artificial intelligence and statistics*, pages 4519–4529. PMLR, 2020.

[LSAS23] Bo Li, Mikkel N Schmidt, Tommy S Alstrøm, and Sebastian U Stich. On the effectiveness of partial variance reduction in federated learning with heterogeneous data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3964–3973, 2023.

[MMB12] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

[MSB12] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[RVRK16] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016.

[SK19] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[SQLG15] Hao Su, Charles R Qi, Yangyan Li, and Leonidas J Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *Proceedings of the IEEE international conference on computer vision*, pages 2686–2694, 2015.

[SZY23] Tingwei Shen, Ganning Zhao, and Suya You. A study on improving realism of synthetic data for machine learning. In *Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications*, volume 12529, pages 270–277. SPIE, 2023.

[Tzi22] Rigas Tzikas. How realistic is my synthetic data? a qualitative approach. Master's thesis, 2022.

[WBL23] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023.

[WCL23] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2555–2563, 2023.

[YAH23] Kun Yuan, Sulaiman A Alghunaim, and Xinmeng Huang. Removing data heterogeneity influence enhances network topology dependence of decentralized sgd. *Journal of Machine Learning Research*, 24(280):1–53, 2023.

[ZIE+18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Acronyms

AI Artificial Intelligence.

ASM Angular Second Moment.

ATD Automated Target Detection.

ATR Automated Target Recognition.

BRISQUE Blind/Referenceless Image Spatial QUality Evaluator.

CLEANSE Checking Learning Entries and Anomalies for Network Stability and Efficiency.

CLIP Constrastive Language-Image Pretraining.

CLIP-IQA Constrastive Language-Image Pretraining Image Quality Assessment.

CNN Convolutional Neural Network.

DL Deep Learning.

EO Electro-Optical.

FaRADAI Frugal And Robust AI for Defence Advanced Intelligence.

FID Frechet Inception Distance.

GLCM Gray Level Co-occurrence Matrix.

IoU Intersection over Union.

IQA Image Quality Assessment.

IR Infrared.

LPIPS Learned Perceptual Image Patch Similarity.

LWIR Long Wave Infrared.

ML Machine Learning.

MODTRAN MODerate resolution atmospheric TRANsmission.

MSCN Mean Subsracted Contrast Normalization.

NIQE Natural Image Quality Evaluator.

NN Neural Network.

RADAR RAdio Detection And Ranging.

RF Radio-Frequency.

RND Random Network Distillation.

SE Spectral Entropy.

SE-WB SE-WORKBENCH.

SE-WB-AI SE-WORKBENCH-AI.

Self-FID Self-FID.

SER Spectral Energy Ratio.

SML Sum of Modified Laplacian.

SNS Scene Natural Statistics.

SVM Support Vector Machine.

SWIR Short Wave Infrared.

YOLO You Only Look Once.