Assessing fidelity in synthetic datasets: A multi-criteria combination methodology

Alexandra Duminil
COSYS/PICS-L
Université Gustave Eiffel
Marne-la-Vallée, France
alexandra.duminil@univ-eiffel.fr

Sio-Song Ieng
COSYS/PICS-L
Université Gustave Eiffel
Marne-la-Vallée, France
sio-song.ieng@univ-eiffel.fr

Dominique Gruyer
COSYS/PICS-L
Université Gustave Eiffel
Marne-la-Vallée, France
dominique.gruyer@univ-eiffel.fr

***Abstract*—With the development of driving simulators, graphics engines and synthetic-to-real domain adaptation algorithms, synthetic datasets become increasingly more photo-realistic. The advancement of such dataset is crucial for advanced driving systems, particularly for training learning-based methods and validation. An important consideration is around the fidelity of synthetic datasets, particularly regarding their suitability for deep learning applications such as object detection or segmentation. However, quantifying fidelity poses a significant challenges. To address this gap, we propose a set of fidelity scores to quantify the level of fidelity of RGB images from these datasets. Through in-depth examination, we aim to reveal information about the texture patterns and high-frequency components that contribute to the objective perception of data realism in road scenes. Furthermore, a multi-criteria combination using belief theory is performed to merge these scores and give a global score involving the level of fidelity, the level of uncertainty on this decision, and the level of conflict between the scores.**

***Index Terms*—fidelity, synthetic dataset, multi-criteria combination, belief theory, automated driving systems, image analysis**

I. Introduction

In the context of functional evaluation and validation of vision- and AI-based automated transport systems, it is necessary to address these steps at both algorithms level and applications used by automated mobility systems, and in terms of tools, software and test models, particularly in simulation environments. Ensuring the safety and reliability of automated driving systems is essential and poses significant challenges. The progress of learning-based algorithms make it possible to carry out ever more complex and more accurate tasks, particularly in detection and tracking. Real-world and simulated datasets contribute to this advancement in such tasks. The advent of synthetic data has made possible to overcome the difficulties of acquiring real-world data. Thus, in a simulated environment, it becomes possible to incorporate various scenes, lighting, weather simulation, and sensors with ever-improving techniques based on rendering simulators, game engines or deep-learning methods. However, gaps last between real and simulated data, which can pose significant generalization problems and can lead learning-based algorithms into error.

*This work is supported by French National Research project PRISSMA (pillar 2 of the French great challenges about AI) and Horizon Europe Augmented_CCAM project

The French Grand Challenge on AI project PRISSMA addresses a part of these objectives which is to develop, among other purposes, a methodology for evaluating and validating systems of systems, AI-based systems, test equipment and datasets used for the training and testing. Our research falls in this context and aims to propose a methodology and metrics for assessing the fidelity of synthetic data, used in the evaluation and validation process of AI-based systems. To address this challenge, we propose a set of scores ensuring that simulated data is sufficiently faithful to reality. It becomes essential to quantify the fidelity of these synthetic datasets, thus enabling confident use in the learning, evaluation, and validation stages of perception AI algorithms.

In this paper, we propose to quantify the level of fidelity of synthetic datasets through a set of scores. The produced scores will provide an indication of whether the datasets are faithful enough with respect to chosen features for evaluating automated vehicle perception. Therefore, we will focus on image features. Specifically, we decided to use texture information, with the grey level co-occurrence matrix method (GLCM) [1], a statistical texture analysis technique, enables the evaluation of image structural properties by examining the spatial relationships between pixels, with a set of 14 Haralick metrics. Morevover, we have integrated the local binary pattern (LBP) method to conduct a more localized texture analysis. Furthermore, to address high-frequency information, the wavelet transform is used. Subsequently, we consolidate these proposed scores using a multi-criteria combination method [2]. This approach considers various sources of information, as well as uncertainties and potential conflicts, ensuring a comprehensive assessment.

II. Related works

Recent works have investigated the simulation-to-reality (S2R) gap across various sensor categories, including camera-based [3], [4], RADAR-based [5], and LiDAR-based [6] approaches for object detection algorithms. The S2R gap approach involves training models using both synthetic and real data. In the study by [3], they compared an environment

simulation software with real-world test drives. This comparison assessed the disparity between simulation and reality using metrics like Precision, Recall, Multi-Object Tracking Accuracy (MOTA), and Precision (MOTP), applied to object lists from both datasets. [4] introduces a domain adaptation technique via Conditional Alignment and Reweighting. This method systematically utilizes target labels to explicitly reduce the gap between simulated and real domains, although it does not provide specific scores or metrics. [5] focuses on assessing the fidelity of various radar model types and their applicability for virtually testing radar-based multi-object tracking with a multi-level testing method. [6] aims to quantify the simulation-to-real domain shift by analyzing point clouds at the target level. This is achieved by comparing real-world and simulate point clouds within the 3D bounding boxes of the targets. However, these works were conducted within the context of a specific application which is object detection and tracking using digital twins. A recent approach [7] presents a technique for evaluating the quality levels of synthetic underwater images. It involves extracting statistical, perceptual, and texture-based measures from a transmission map, and suggests incorporating color features and fractal-based texture features. While their work focuses on evaluating the fidelity of synthetic underwater images, our research aims to quantify the fidelity level in computer-generated images.

III. Fidelity scores

A. Definition

A comprehensive conceptual framework of fidelity has already been proposed [8], and can be dividing objective and subjective fidelity. The objective fidelity is the concept that enables a quantification through physical metrics. In this work, fidelity denotes the similarity between selected features in the virtual environment and their corresponding reference features in the real environment. Thus, high fidelity implies a faithful representation, while low fidelity suggests an incomplete and simpler representation.

B. Methodology

A feature-based analysis is conducted in order to study and quantify the level of fidelity of several synthetic datasets. Based on Haralick's assumption [1] and regarding the fundamental pattern present in real images, we chose to analyse the texture and frequency features of images with two complementary approaches. The first approach involves an image classification method that uses GLCM, the local binary pattern [9] (LBP) and wavelet transforms as pre-processing techniques [10] for input data obtained from different synthetic datasets. These pre-processed data are then fed into a CNN network to enable the classification of images as faithful to reality or not. Using feature learning, instead of working directly with raw data, allow generating more interpretable and insightful data representations. Indeed, the main hypothesis is that the fidelity calculation is directly dependant of the features. Hence, different types of features must be considered in order to obtain an accurate measurement of a score. A diagram is presented Fig. 1.

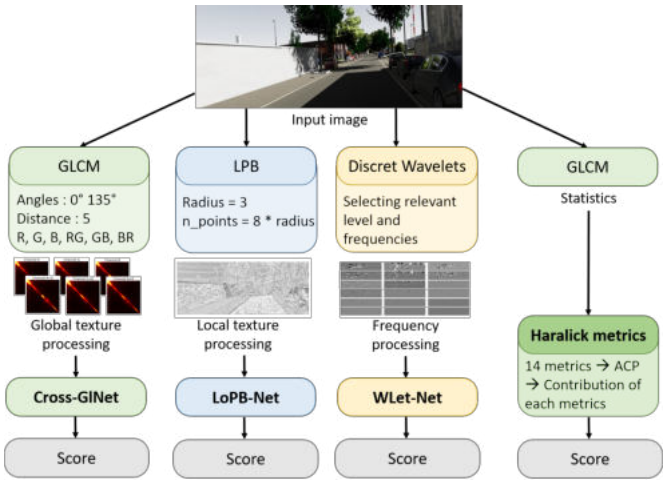

Fig. 1. Diagram of the method.

1) CNN-based methods: We propose three sub-networks, namely Cross-GlNet, WLet-Net and LoPB-Net, trained separately in a supervised manner using a custom dataset. Each of these sub-networks takes respectively GLCM maps, wavelet transforms, and LPB maps calculated from RGB images as inputs into CNN networks.

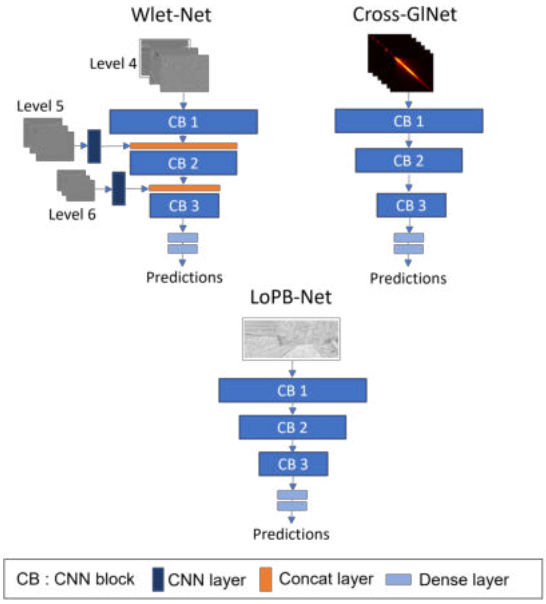

Fig. 2. Schematic structures of the networks.

Inspired by a recent approach [11], the Cross-GlNet model uses the GLCM or LPB maps as input, but there are some differences in their approach. For more clarity, we name CrossGlNet for the CNN with GLCM as input and LoPB-Net for the CNN with LPB as input. Cross-GlNet takes the computation of GLCM in two directions (horizontal and diagonal) and a pixel distance of 5 on the cross-bands RGB

channels (R+G, G+B, B+R) of the images. These GLCMs are then stacked together to form an input tensor with a size of $256 \times 256 \times 6$. We opted to use only cross-band channels because additional experiments have indicated that these inputs yield better performance compared to single RGB channels. The models discussed in this section share a nearly identical architecture, as observed in Fig. 2. They consist of :

- CB 1: A convolutional layer with 32 filters of size 3×3, a batch normalization and ReLu activation followed by a max-pooling layer
- CB 2: A convolutional layer with 64 filters of size 3×3, a batch normalization and ReLu activation followed by a max-pooling layer
- CB 3: A convolutional layer with 128 filters of size 3×3, a batch normalization and ReLu activation followed by a max-pooling layer
- A dense layer with 256 nodes followed by a ReLu layer
- A dense layer with 1 node followed by a Sigmoid layer

The key difference between WLet-Net and the two others (on the right), is the incorporation of multi-scale inputs. WLet-Net uses wavelet transforms that can decompose the resulting image into a combination of levels. As levels 5 and 6 have lower resolutions and finer frequencies than level 4, they are incorporated at a higher stage in the network, fading into a CNN layer with respectively 32 and 64 filters of size 1x1 followed by a concatenation layer. This type of coarse-to-fine architecture has already been proposed [12] and enables to restore high-frequency information through the network. Moreover, incorporating them at higher stages seems appropriate since these inputs are noisier compared to those of level 4. For each model, we utilize the Keras/TensorFlow framework, employing the SGD optimizer with a learning rate of 0.0001 and using binary cross-entropy as the loss function. Batch size is set to 32, and epochs start at 40, with an early stopping to reduce the risk of overfitting.

For the learning based methods, a custom dataset is used and brings together the data from the datasets mentioned in the following. Several well-known and publicly available datasets have been collected that are either real or simulated. The real datasets include KITTI [13], Cityscapes [14], ONCE [15] and NuScenes [16] and the synthetic datasets consist of virtual KITTI (vKitti) [17], KITTI-CARLA [18], GTA V [19] and Synthia [20]. In our methodology, leveraging both real and synthetic images depicting nearly identical scenes is both interesting and desirable. This approach facilitates a more effective and dependable comparison and interpretation. This custom dataset was split into 3 subsets for the algorithm training, the algorithm validation and the testing: There are 20572 images in the training set, 6755 in the validation set and 1000 in the test set.

2) Haralick metrics: The use of Haralick metrics provides a different approach compared to the models usually employed for various tasks. It make possible to compare information acquired in simulated images, thereby determining how realistic there are. The Haralick metrics used in this paper are : angular second moment (ASM), contrast, correlation, Sum of squares : variance (variance), inverse difference moment (IDM), sum average, sum entropy, entropy, difference variance, difference entropy, info measure of correlation 1 (IMC1), info measure of correlation 2 (IMC2).

The Haralick metrics are calculated on four synthetic and real datasets (mean on images per dataset), with a total of 100000 image patches of size 64 by 64 pixels. To highlight the most relevant metrics by image type, we propose to directly applying the Principal Component Analysis (PCA). PCA is generally used to reduce the dimensionality of data, but it is also an effective tool for analyzing and interpreting. This approach is employing to better understand the individual contribution of each metrics to the overall information, the links between the various metrics and their contribution on each principal component (PC). As the first two components contain over 50 % of the data's information, we focus on the first two PCs for dataset analysis. Then, we compute the contribution of each metric to each principal component PC_1 and PC_2 with:

$$K_{i,k} = \frac{c_{i,k}^2}{\lambda_k} \tag{1}$$

where k is the PC index, i is the metric index, λ_k is the eigenvalue associated to the PC_k and $c_{i,k}$ is the component of the vector $\sqrt{\lambda_k}\mathbf{u_k}$ for the i^{th} metric and $\mathbf{u_k}$ is the k^{th} eigen vector.

TABLE I
CONTRIBUTION OF EACH METRIC TO PC_2. THE BEST CONTRIBUTIONS AMONG THE SYNTHETIC DATASETS ARE IN BOLD. THE BEST CONTRIBUTIONS AMONG THE REAL DATASETS ARE UNDERLINED.

Metrics	Kitti	City	Once	NuScenes	vKitti	GTAV	Kitti-C	Synthia
ASM	0.015	0.15	0.095	0.032	0.027	0.040	0.015	0.091
Contrast	0.064	0.10	0.19	0.16	0.009	0.029	0.071	0.020
Corr	0.055	0.016	0.030	0.006	**0.18**	**0.15**	**0.14**	**0.10**
Var	0.18	0.19	0.21	0.27	0.008	0.11	0.036	0.079
IDM	0.033	0.12	0.012	0.024	0.006	0.074	0.12	0.13
SA	0.018	0.005	0.079	0.084	**0.34**	0.009	**0.25**	**0.16**
SVar	0.22	0.19	0.21	0.28	0.013	0.12	0.028	0.084
SE	0.002	0.011	0.025	0.019	0.0005	0.001	2e-5	0.009
E	0.014	0.031	0.026	0.026	4e-5	0.019	0.003	0.036
DVar	0.077	0.088	0.12	0.075	0.056	0.008	0.081	0.005
DE	0.12	0.12	0.13	0.13	0.003	0.020	0.049	0.053
IMC1	0.28	0.10	0.005	0.06	**0.17**	**0.30**	**0.13**	**0.18**
IMC2	0.037	0.002	0.084	0.014	0.18	0.11	0.076	0.042

We will focus on presenting the results from PC_2, as it is the most informative and serves as the primary basis for our analysis. Table I presents the computed contribution of each metric to PC_2 for different datasets. In Table I, we can observe that some metrics are more significant for the synthetic datasets (correlation, sum average, and IMC1), while others (var and svar) are more indicative for the real datasets. These results will be used to create a fidelity score.

C. Resulting fidelity scores

Several indices need to be considered when quantifying the image fidelity due to the complexity of real scenes. Indeed, it

is not sufficient to only use Haralick metrics for establishing an accurate score of fidelity. Therefore, we propose a set of scores including models and the selected Haralick metrics, that provides a more comprehensive assessment of fidelity. The last sub-score is defined by the following equation :

$$sH = \frac{1}{5}(\lambda_2 K_{Var,2} + \lambda_2 K_{Svar,2} + (1 - \lambda_2 K_{Corr,2}) + (1 - \lambda_2 K_{SA,2}) + (1 - \lambda_2 K_{IMC1,2})) \quad (2)$$

Equation 2 is using arithmetic average of the correlation contributions $\lambda_2 K_{i,2}$ of selected metrics. It takes into account all the best contribution to PC_2 for both synthetic and real datasets. Table II presents the fidelity scores of synthetic datasets employing the different proposed methods, including learning-based methods and Haralick metrics. The fidelity scores generated from the models (GLCM, Wavelets, LPB) correspond to the prediction functions acquired through the Keras/Tensorflow frameworks.

TABLE II
FINAL FIDELITY SCORES (%).

Methods	vKitti	Kitti-C	Synthia	GTA V	GTA/Map
GLCM	0.05	0.37	10.42	12.03	21.04
Wavelets	51.21	1.61	3.60	4.30	29.92
LPB	10.44	4.03	13.54	11.58	36.88
sH	34.10	38.98	41.80	48.14	81.89

GTA/Map is a GAN-enhanced version of GTA dataset [21] which exhibit higher fidelity scores. Nonetheless, the evaluated synthetic datasets generally demonstrate low fidelity. To enhance the representation metric, we propose using a multi-criteria combination operator. This approach merges scores and quantifies uncertainty and conflict in the obtained results.

IV. MULTI-CRITERIA COMBINATION RULES

A. Introduction on Belief Function Theory

Multi-criteria combination based on belief function theory [2] makes it possible to merge information from several sources (in our case, the scores) that may be independent, facilitating multi-criteria analysis and decision-making processes. Moreover, it effectively handles uncertainty and conflicts between different criteria. This method is based on the Belief Function Theory (or Dempster-Shafer Theory) which is often used as a sensor fusion method [22]. It allows to obtain degrees of belief for a particular inquiry from subjective probabilities for a related hypothesis, denoted H. Initially, this involves a combination of several assumptions. However, in our case, the process is simplified as we have only one hypothesis. Let us define Ω as the frame of discernment, which operate under close world assumptions.

These assumptions, which can be true or false, are quantifying using a mass written $m(H)$ and defined as :

$$m^{\Omega} : 2^{\Omega} \rightarrow [0, 1] \quad (3)$$

Then, if $m(H) = 0$, we have no evidence supporting the hypothesis, and if $m(H) = 1$, it is supposed that we have all evidence supporting the hypothesis. A source of information gives its opinion on the truth of hypotheses. The set of masses provided by this source is called the Basic Belief Assignment (BBA). Based on the Dempster-Shafer theory, the sum of the mass associated with BBA must be equal to 1.

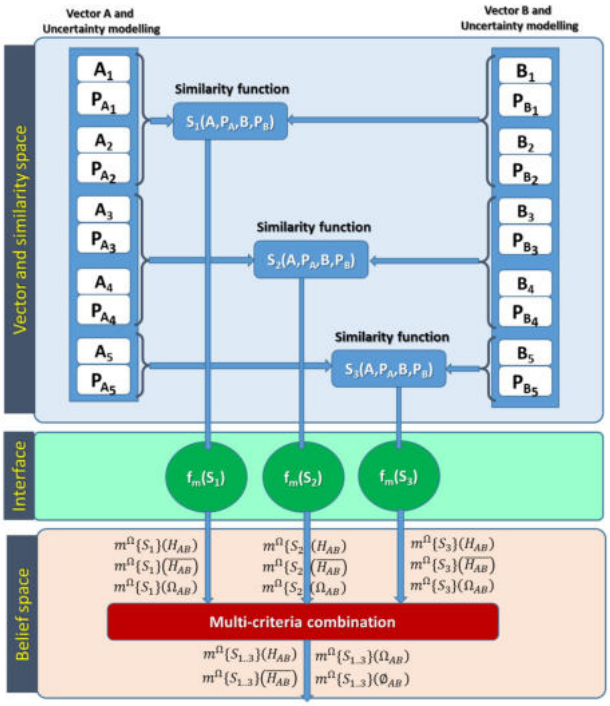

Fig. 3. Overview of the multi-criteria combination method with the different stages (with 3 criteria and 5 elements state vectors).

A similarity index S_k is calculated for each criterion k, which is converted into a BBA by using a set of Basic Belief Functions (BBF) making the interface between initial space (with units) represented by a similarity or dissimilarity metric toward the symbolic belief space. In this context, the BBA named S_k have the following form :

$$S_k = \{m^{\Omega}\{Sc_k\}(H), m^{\Omega}\{Sc_k\}(\bar{H}), m^{\Omega}\{Sc_k\}(\Omega)\} \quad (4)$$

The different mass-sets have to be calculated for each criterion, thanks to a mass-generative function BBF, in order to combine them as a multi-criteria combination operator. A mass generative function must respect the following form :

$$f_m : \Re \rightarrow \Re^3 \quad (5)$$

$$Sc_k \mapsto S_k \quad (6)$$

where Sc_k corresponds to the similarity indices coming from the score k such as $Sc_k \in [0; 1]$ and S_k is a set of masses which have to satisfy the following requirement:

$$m^{\Omega}\{Sc_k\}(H) + m^{\Omega}\{Sc_k\}(\bar{H}) + m^{\Omega}\{Sc_k\}(\Omega) = 1 \quad (7)$$

For each score (information source about the fidelity), the value is bounded to the interval [0, 1], the maximum score being equal to 1 (the image is considered as similar to actual data).

In Fig. 3, an example is given with the multi-criteria combination of 2 vectors with 5 components with their associated uncertainties. In order to generate the 3 criteria, a set of similarity functions are used (for instance Gruyer's, Mahalanobis, or Bhattacharyya distances). The first criterion is made from the use of the 2 first elements of the 2 state vectors A, B and the associated uncertainty P_A and P_B. The second criterion uses the third and fourth elements. The last criterion only uses the similarity between the last element of A and B vectors. This example is interesting because it shows the possibilities offered by this methodology. Indeed, a state vector can be used with sub group of vector's component. This operating could be useful depending on the type of modelling we will used to represent the uncertainty, or the type of data we will handle (data heterogeneity). Moreover, depending on the data availability (different operating frequencies for a sub part of the vector elements), it is always possible to generate a set of masses for each criterion. If, for a criterion, we don't have enough information in the vectors to calculate the similarity, then the BBF will generate the following set of masses: $m_k(H_{AB})$ = $m_k(\bar{H}_{AB})$ = 0.0, and $m_k(\Omega_{AB})$ = 1.0; This set of masses means we have a total ignorance about the current situation. In this condition, this method could operate with partial, asynchronous, and heterogeneous data.

This mass-set is used in the multi-criteria combination rule which consists in computing a synthetic mass-set from the combination of K mass-sets provided by K sources of information. The multi-criteria combination rule is based on generalization of Dempster-Shafer's (D-S) conjunctive combination rule, which is as follow, for two sources of information S_1 and S_2, and for any hypotheses H_X, H_Y, and H_Z:

$$m_{1,2}(H_X) = \frac{1}{1-K} * \sum_{H_Y \cap H_Z = H_X} m_1(H_Y).m_2(H_Z) \quad (8)$$

Then, a recursive aggregation of these sources is performed, to leads to the generalized K-source combination formula. The recursive formulation of the D-S combination rules can be formulated as follow:

$$m_{1..k+1}(H_X) = \frac{1}{1-K} * \sum_{H_Y \cap H_Z = H_X} m_{1.k}(H_Y).m_{1.k+1}(H_Z) \quad (9)$$

It integrates $k+1$ sources of information, considering the result of the previous combination and combined with the current source. Indices behind the m represent the sources involved in the mass calculation. The combination of sources $\{H, \bar{H}, \Omega\}$ results on four hypotheses, which are : $\{\Theta, H, \bar{H}, \Omega\}$. Θ represents conflict between sources.

To better understand the various hypotheses, let's consider the hypothesis where H_{AB} corresponds to : the vectors A and B are identical. This scenario presents four potential mass values:

- $m_{1..K(H_{AB})}$: Mass about the similarity between the vectors A and B.
- $m_{1..K(\bar{H}_{AB})}$: Mass about the dissimilarity between the vectors A and B.
- $m_{1..K(\Omega_{AB})}$: Mass representing the uncertainty between A and B. We don't know if A and B are similar or not.
- $m_{1..K(\Theta_{AB})}$: Mass on the conflict, or non-accordance between the various criteria. This means a part of the vector A is similar to B and the rest is dissimilar.

B. Mass-sets generation for scores combination

In our context, the objective is not to generate a similarity index with belief function theory, as mentioned in [2] and addressed in the previous section, but to adapt this method in order to combine a set of scores Sc_k obtained from 4 previous pre-processing functions. Each score initially takes a value in $[0..100]$ and need to be translated in the space $[0..1]$. Moreover, the frame of discernment is built with only one hypothesis H represented the assertion *is faithful to the reality produced by a camera*. Then, the objective consists to model the set of scores Sc_k as similarity indices and to convert them into BBA by using a set of BBF making the interface between initial score space and the symbolic belief space.

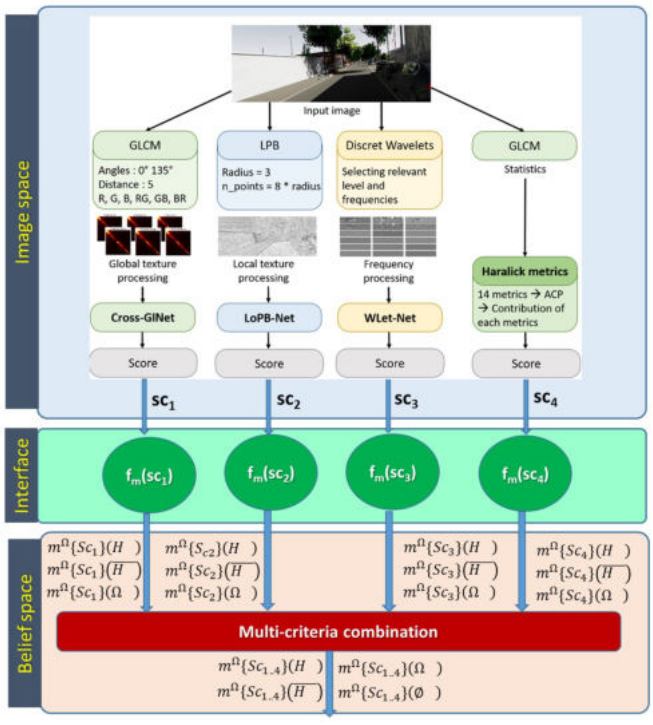

Fig. 4. Overview of the multi-criteria combination method for the assessment of a global fidelity score involving uncertainty and potential conflict detection.

From the score sc_k coming from one of the 4 evaluation methods, a BBA is computed on the following postulate: *An information source cannot generate belief into two contradictory propositions*. In fact, a score S_k cannot, at the same time,

assert that an evaluated image is and is not realistic (from the fidelity point of view). This assumption is essential because it guarantees a physical and semantic consistence. However, we can obtain, at the same time, a mass on H and Ω, or on $\bar{H}$ and Ω. This means that the current score quantify a positive opinion about the fidelity of the data but with a doubt (level of uncertainty). It also means that a BBA can not be generated with both focal elements H and $\bar{H}$. Indeed, considering that information sources are reliable, physical meaning leads to define the BBA by limiting the mass of conflict as follows:

$$\begin{aligned} m^{\Omega}\{Sc_k\}(H) &= \begin{cases} 0 & Sc_k \in [0,\tau_2] \\ \Phi_1(\alpha_0, Sc_k) & Sc_k \in [\tau_2,1] \end{cases} \\ m^{\Omega}\{Sc_k\}(\bar{H}) &= \begin{cases} 0 & Sc_k \in [\tau_1,1] \\ \Phi_2(\alpha_0, Sc_k) & Sc_k \in [0,\tau_1] \end{cases} \\ m^{\Omega}\{Sc_k\}(\Omega) &= \begin{cases} 1-\Phi_2(\alpha_0, Sc_k) & Sc_k \in [0,\tau_1] \\ 1 & Sc_k \in [\tau_1,\tau_2] \\ 1-\Phi_1(\alpha_0, Sc_k) & Sc_k \in [\tau_2,1] \end{cases} \end{aligned} \tag{10}$$

where $\Phi_i, i \in \{1,2\}$ are BBF and τ_i a threshold sharing the functions space. In this manner, the coefficient α_0 can be seen as a specific coefficient discounting each source, as proposed by Appriou in [23] with the definition of specialized sources.

In our application, the α_0 coefficient can be seen as a reliability coefficient concerning the method provided a specific score Sc_k (see Fig. 5). Considering this remark, α_0 is therefore considered as constant over time and for any image processed by a method (GLCM, Wavelets, LPB, and sH). In a generic framework, both functions Φ_1 and Φ_2 must be chosen or built as bijective functions and considering the following limiting conditions:

$$\begin{cases} \Phi_1(\alpha_0, 1) = \alpha_0 \\ \Phi_1(\alpha_0, \tau_2) = 0 \end{cases} \tag{11}$$

$$\begin{cases} \Phi_2(\alpha_0, 0) = \alpha_0 \\ \Phi_2(\alpha_0, \tau_1) = 0 \end{cases} \tag{12}$$

Numerous bijective functions can be used, from simple to elaborated ones [24]. The most simple model consists to generate BBA with BBF defined using two linear functions: a function for each interval $[0,\tau_1]$ and $[\tau_2,1]$. The linear functions have a slope of $-\frac{\alpha_0}{\tau_1}$ and $\frac{\alpha_0}{1-\tau_2}$ respectively. These parameters settle between the fact that the source of information is either in favor or against the association.

Fig. 5 and 6 are a specific case with $\tau_1 = \tau_2 = \tau$. When $\tau = 0.5$, the system is said to be *neutral*. When τ is lower than 0.5 then the system is said to be *optimistic* because the model tends to allocate mass on H even if the similarity Sc_k is weak. When τ is bigger than 0.5, then the system is said to be *pessimistic*, because the similarity should be high in order to allocate mass on H. In our application, we use this specific one-coefficient transformation to compute the BBA. However, more elaborated functions than linear functions can be used for Φ_1 and Φ_2.

Below are the mass-generative functions Φ_1 and Φ_2 ($f_m(sc_k)$)(13)(14) used in this paper. These functions are defined using cosinus functions as follows and as illustrated in Fig. 5:

$$\Phi_1(\alpha_0, \tau, Sc_k) = \frac{\alpha_0}{2}\left(1 - \cos\left(\pi.\frac{Sc_k - \tau}{1-\tau}\right)\right) \tag{13}$$

$$\Phi_2(\alpha_0, \tau, Sc_k) = \frac{\alpha_0}{2}\left(1 + \cos\left(\pi.\frac{Sc_k - \tau}{1-\tau}\right)\right) \tag{14}$$

The choice of such functions is justified by their properties of continuity and derivability on the intervals. However this choice is not subject to a detailed study.

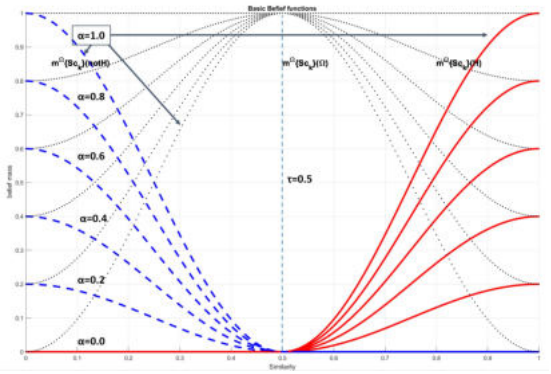

Fig. 5. Generation of BBA with BBF with a variation of α and $\tau = 0.5$.

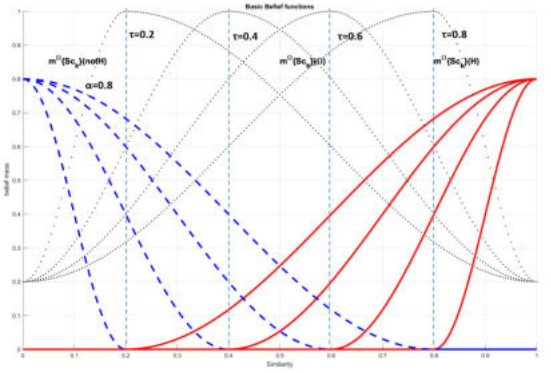

Fig. 6. Generation of BBA with BBF with an $\alpha = 0.8$ and a variation of τ.

C. Multi-criteria combination rules

The sources are expressed on the same triplet of hypotheses denoted by $H, \overline{H}, \Omega$ characterizing the level of fidelity of a synthetic or real image depending on a specific processing. Thus, we have a frame of discernment common to all sources allowing to apply the generalized conjunctive combination operator ([2]) on a common hypothesis (also called the multi-criteria combination operator).

$$\begin{aligned} m_{Sc_1..Sc_N}(H) = \prod_{j=1..N}(1 - m^{\Omega}\{Sc_j\}(\bar{H})) - \\ \prod_{j=1..N}(m^{\Omega}\{Sc_j\}(\Omega)) \end{aligned} \tag{15}$$

$$m_{Sc_1..Sc_N}(\overline{H}) = \prod_{j=1..N}(1 - m^{\Omega}\{Sc_j\}(H)) - \prod_{j=1..N}(m^{\Omega}\{Sc_j\}(\Omega)) \quad (16)$$

$$m_{Sc_1..Sc_N}(\Omega) = \prod_{j=1..N}(m^{\Omega}\{Sc_j\}(\Omega)) \quad (17)$$

$$m_{Sc_1..Sc_N}(\emptyset) = 1 - \prod_{j=1..N}(1 - m^{\Omega}\{Sc_j\}(\bar{H})) - \prod_{j=1..N}(1 - m^{\Omega}\{Sc_j\}(H)) + \prod_{j=1..N}(m^{\Omega}\{S_j\}(\Omega)) \quad (18)$$

D. Results

In this subsection, the multi-criteria combination method is applied to the fidelity scores, presented in section III. This analysis is conducted across three datasets: the synthetic GTA V dataset, the GAN-enhanced version of GTA dataset GTA/Map, and the real Cityscapes dataset. GTA/Map is expected to be more faithful than GTA V, as seen in Table II. Hence, it is pertinent to examine the impact of the results using the multi-criteria approach. The Cityscapes dataset serves here as a reference. Four scores have been calculated for each dataset, corresponding to the four criteria, as shown in Fig. 4.

Fig. 7 illustrates the graphs obtained from the multi-criteria combination and the generation of BBA with BBF. They will help to establish a level of fidelity (H) or non fidelity ($notH$), the level of uncertainty ($Omega$), and the detection of conflict ($Empty$) between scores. Table III details the parameters used to produce these graphs. Each criterion has been assigned reliabilities α and τ values. The τ values are set here to be pessimistic with τ = 0.6 (model tends to allocate mass on $\overline{H}$). For the time being, τ is fixed, but it will be optimized as part of a future work. The reliabilities associated with the criteria, obtained from the learning-based models, correspond to models' accuracy. These accuracies are computed using functions from the the Keras/TensorFlow framework. The reliability assigned to the S_H criterion is set to 0.5 as it is impossible to obtained a similar accuracy to the learning-based methods.

TABLE III
RELIABILITY α AND τ ASSOCIATED TO EACH CRITERION SC WITH A CERTAIN VALUE FOR THREE DATASETS.

Criteria	GTA V			GTA/Map			Cityscapes		
	value	α	τ	value	α	τ	value	α	τ
Sc_1	0.12	0.92	0.6	0.21	0.64	0.6	0.97	0.99	0.6
Sc_2	0.04	0.97	0.6	0.30	0.73	0.6	0.98	0.99	0.6
Sc_3	0.11	0.90	0.6	0.37	0.70	0.6	0.99	0.99	0.6
Sc_4	0.48	0.50	0.6	0.82	0.50	0.6	0.72	0.50	0.6

The graph on the left in Fig. 7, for the GTA dataset, indicates that it has a strong tendency to $\overline{H}$ with 97%, as well as 2% of conflicts. These results suggest that the GTA V dataset is not faithful to reality. The graph on the right, resulting from

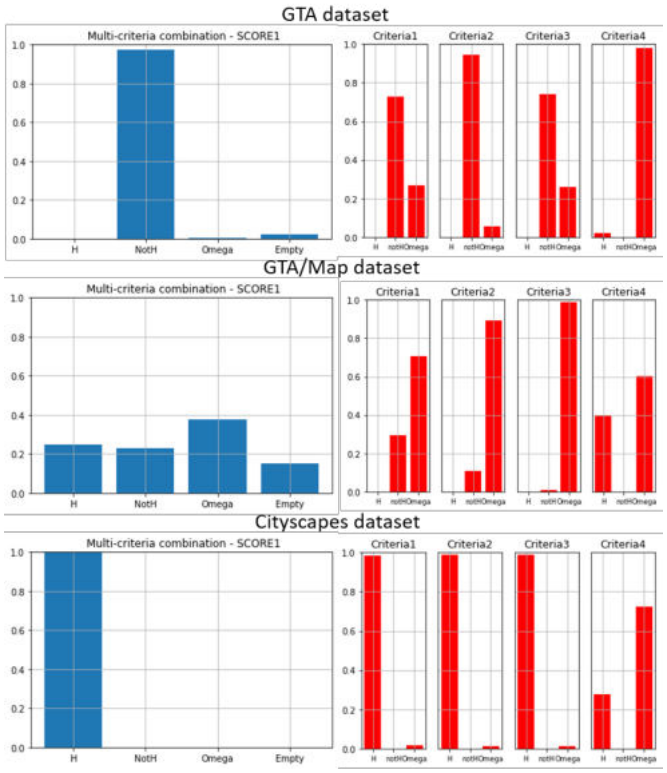

Fig. 7. Graphs resulting from the multi-criteria combination (left) and the generation of BBA with BBF (right).

the generation of BBA, present the outcomes concerning the triplet of hypotheses $H, \overline{H}, \Omega$ for each criterion. This allows us to obtain detailed results for each criterion with $m(H)$ = 0%, 0%, 0%, 2%, $m(\overline{H})$ = 73%,95%, 74%, 0%, $m(\Omega)$ = 27%, 5%, 26%, 98%. We can see that criterion 4 shows a tendency of 2% to H and an uncertainty of 98%.

Concerning the GTA/Map dataset, it indicates a weaker tendency to $\overline{H}$ than the GTA V dataset with 23%. It also includes 25% of H, 38% of uncertainty and 15% of conflicts. Concerning the triplet of hypotheses : $m(H)$ = 0%, 0%, 0%, 40%, $m(\overline{H})$ = 29%, 11%, 1%, 0%, $m(\Omega)$ = 70%, 89%, 99%, 60%. While the level of uncertainty is very high for all criteria, the first shows a significant tendency to $\overline{H}$ and the fourth an increasing tendency to H with 40% and an uncertainty of 60 %. This shift of $\overline{H}$, which correspond to a level shift of non fidelity from 97% to 24%, suggests an improved fidelity of the enhanced GTA dataset compared to the GTA V dataset. The third row of the figure shows the graphs for the real Cityscapes dataset. These results serve as an ideal basis for datasets requiring to be faithful to real-world scenarios.

V. CONCLUSION

The advancement of automated mobility necessitates rigorous evaluation and validation of its components, particularly those driven by AI systems. Creating representative datasets for training is challenging, prompting the use of simulation methods. However, assessing the fidelity of synthetic road

images remains a significant challenge. Our research aims to evaluate virtual images comprehensively using a distinct set of metrics, addressing the lack of objective assessment methods in existing literature. By employing metrics like GLCM, wavelet transforms, LPB, and Haralick-based ([25]), we can quantify texture information effectively. We found that combining statistics-based and learning-based metrics offers valuable insights into dataset fidelity. Our approach leverages the strengths of model-based and statistical methods, providing versatility across various applications. We propose a set of metrics to quantify the fidelity of synthetic images, crucial for determining their suitability for training AI-based perception systems. Then from the 4 metrics providing fidelity scores, we propose an innovative combination of these scores by using belief function theory in order to generate a final and global score. This global score provides 4 interesting information about the fidelity of a synthetic dataset: the level of fidelity, the level of non fidelity, the level of uncertainty about this decision, and the detection of conflict between local scores. This work lays the foundation for efficient and objective labeling and certification of virtual data.

Future research will delve deeper into analyzing different types of road scenes to gather more accurate information. Exploring different color spaces before applying feature extraction techniques could optimize our method further. Additionally, studying fractal-based features can offer insights into image structural complexity. Understanding the characteristics distinguishing realistic from synthetic images is crucial, potentially informing models of *real data* and guiding modifications to virtual data for increased realism. This prompts the question of specific attributes contributing to the perception of realism from a camera perspective. Two last improvements in progress are related to belief theory. A first one consist to apply an optimization process allowing to tune the hyper parameters like the reliability α_i and τ border of BBF in order to minimize the conflict or the level of uncertainty. A second one reuse the figure 3 with the vector A containing the features of the synthetic image and the vector B the features providing the model of a realistic image. In this context, the multi-criteria combination will allow to identify the gap between the synthetic data and the model. This gap with the identification of the features generating it (using the conflict) will allow to provide advice for the synthetic data improvements.

REFERENCES

[1] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[2] V. Magnier, D. Gruyer, and J. Godelle, "Multi-criteria similarity operator based on the belief theory: Management of similarity, dissimilarity, conflict and ambiguities," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 1215–1221.

[3] F. Reway, A. Hoffmann, D. Wachtel, W. Huber, A. Knoll, and E. Ribeiro, "Test method for measuring the simulation-to-reality gap of camera-based object detection algorithms for autonomous driving," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 1249–1256.

[4] V. Prabhu, D. Acuna, A. Liao, R. Mahmood, M. T. Law, J. Hoffman, S. Fidler, and J. Lucas, "Bridging the sim2real gap with care: Supervised detection adaptation with conditional alignment and reweighting," *arXiv preprint arXiv:2302.04832*, 2023.

[5] A. Ngo, M. P. Bauer, and M. Resch, "A multi-layered approach for measuring the simulation-to-reality gap of radar perception for autonomous driving," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 4008–4014.

[6] S. Huch, L. Scalerandi, E. Rivera, and M. Lienkamp, "Quantifying the lidar sim-to-real domain shift: A detailed investigation using object detectors and analyzing point clouds at target-level," *IEEE Transactions on Intelligent Vehicles*, 2023.

[7] X. Li, H. Xu, G. Jiang, M. Yu, T. Luo, X. Zhang, and H. Ying, "Underwater image quality assessment from synthetic to real-world: Dataset and objective method," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 3, pp. 1–23, 2023.

[8] X. Ye, P. Backlund, J. Ding, and H. Ning, "Fidelity in simulation-based serious games," *IEEE Transactions on Learning Technologies*, vol. 13, no. 2, pp. 340–353, 2019.

[9] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[10] K. H. Ghazali, M. F. Mansor, M. M. Mustafa, and A. Hussain, "Feature extraction technique using discrete wavelet transform for image classification," in *2007 5th Student Conference on Research and Development*. IEEE, 2007, pp. 1–4.

[11] M. Barni, K. Kallas, E. Nowroozi, and B. Tondi, "Cnn detection of gan-generated face images based on cross-band co-occurrences analysis," in *2020 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2020, pp. 1–6.

[12] I. Lelekas, N. Tomen, S. L. Pintea, and J. C. van Gemert, "Top-down networks: A coarse-to-fine reimagination of cnns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 752–753.

[13] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[14] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] J. Mao, M. Niu, C. Jiang, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, J. Yu, C. Xu, *et al.*, "One million scenes for autonomous driving: Once dataset," 2021.

[16] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[17] Y. Cabon, N. Murray, and M. Humenberger, "Virtual kitti 2," 2020.

[18] J.-E. Deschaud, "Kitti-carla: a kitti-like dataset generated by carla simulator," *arXiv preprint arXiv:2109.00892*, 2021.

[19] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 102–118.

[20] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3234–3243.

[21] S. R. Richter, H. A. AlHaija, and V. Koltun, "Enhancing photorealism enhancement," *arXiv:2105.04619*, 2021.

[22] A. P. Dempster, "A generalization of bayesian inference," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 30, no. 2, pp. 205–232, 1968.

[23] A. Appriou, "Situation assessment based on spatially ambiguous multisensor measurements," *International journal of intelligent systems*, vol. 16, no. 10, pp. 1135–1166, 2001.

[24] D. Gruyer and E. Pollard, "Credibilistic imm likelihood updating applied to outdoor vehicle robust ego-localization," in *Proceedings of the 14th International Conference on Information Fusion (Fusion 2011), 5-8 July 2011, Chicago, USA.*, 2011.

[25] T. Lofstedt, P. Brynolfsson, T. Asklund, T. Nyholm, and A. Garpebring, "Gray-level invariant haralick texture features," *PLoS ONE*, vol. 14, February 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6386443/pdf/pone.0212110.pdf