Multi-Agent Cooperative Camera-Based Semantic Grid Generation

Antoine Caillot, Safa Ouerghi, Yohan Dupuis, Pascal Vasseur, Rémi Boutteau

Journal of Intelligent & Robotic Systems (2024) 110:64
https://doi.org/10.1007/s10846-024-02093-4

REGULAR PAPER

Multi-Agent Cooperative Camera-Based Semantic Grid Generation

Antoine Caillot[1] · Safa Ouerghi[1] · Yohan Dupuis[2] · Pascal Vasseur[3] · Rémi Boutteau[4]

Received: 14 February 2023 / Accepted: 22 March 2024 / Published online: 22 April 2024

Abstract
The idea of cooperative perception for navigation assistance was introduced about a decade ago with the aim to increase safety on dangerous areas like intersections. In this context, roadside infrastructure appeared very recently to provide a new point of view of the scene. In this paper, we propose to combine the Vehicle-To-Vehicle (V2V) and Vehicle-To-Infrastructure (V2I) approaches in order to take advantage of the elevated points of view offered by the infrastructure and the in-scene points of view offered by the vehicles to build a semantic grid map of the moving elements in the scene. To create this map, we chose to use camera information and 2-Dimentional (2D) bounding boxes in order to minimize the impact on the network and ignored possible depth information as opposed to all state-of-the art methods. We propose a framework based on two fusion methods: one based on the Bayesian theory and the other on the Dempster-Shafer Theory (DST) to merge the information and chose a label for each cell of the semantic grid in order to assess the best fusion method. Finally, we evaluate our approach on a set of datasets that we generated from the CARLA simulator varying the proportion of Connected Vehicle (CV) and the traffic density. We also show the superiority of the method based on the DST with a gain on the mean intersection over union between the two methods of up to 23.35%.

Keywords Intelligent transportation systems · Cooperative mapping · Vehicle-to-everything · Dempster-Shafer theory

1 Introduction

Enhancing safety stands as a pivotal priority within the transportation sector. One of its paramount challenges involves mitigating unforeseen and unpredictable circumstances [1]. A viable strategy entails leveraging the collaborative efforts of both Road Side Unit (RSU) and road users to aggregate extensive data, enabling comprehensive situational awareness and facilitating the anticipation of potential hazards. Nonetheless, the collection of data from various point of view (PoV) may engender conflicting observations necessitating adept management strategies.

In this paper, we propose a method using the in-the-scene PoV of the vehicles along with the infrastructure perception to build a sematic grid map and improve its accuracy. Since we are focusing our approach on the issues of map construction and merging from multiple PoV, the study of the impact of sensor position and synchronization noise is beyond the scope of this paper. However, thanks to the GPS time-stamping and our choice to use considerably reduced data size, we believe that synchronization is not a real problem. The introduced method of cooperative semantic grid map generation is solely based on vision [1]. The presented solution contrasts with the state-of-the-art methods which are either single-view based or obtain depth information. Bayes theory-based and DST-based merging methods are compared and evaluated on a new cooperative dataset[2].

This article is an evolution of our previous work, aiming at generating occupancy grids from multiple agents' PoV [2]. In the remainder of this section, we presents the related works. In Section 2, we introduce the global architecture of our approach. In the next section, we present the methods for creating local semantic grids. The fusion of semantic grids is presented in Section 4 and the decision-making in Section 5. Finally, in Section 6, we discuss the performance of our approach before concluding in Section 7.

Antoine Caillot, Safa Ouerghi, Yohan Dupuis, Pascal Vasseur and Rémi Boutteau contributed equally to this work.

✉ Antoine Caillot
antoine.caillot@aist.go.jp

Extended author information available on the last page of the article

[1] https://github.com/caillotantoine/Coop-Evidential-Semantic-Grid

[2] https://github.com/caillotantoine/carla-V2X-dataset-generator

1.1 Related Works

Occupancy grids have been used for many decades in mobile robotics and are an effective way to represent navigable areas in an environment [3, 4]. It is certainly due to its simplicity of design and understanding that this mapping method is still very popular in the community today [5].

To generate these maps, the original approach consists in equipping a robot with a range sensor and observing the distances to the obstacles forming the environment around the robot [3, 4, 6]. With the formalization of occupancy grid merging methods, it became possible to generate more complex occupancy grids using the displacements within the environment [7]. Another approach to generate more complete occupancy grids is to use multiple point of view (PoV) [8]. This approach was then implemented in an automotive context [5, 9]. The occupancy grid concept can be extended to show other information such as the semantics of a cell [5, 10–12] forming semantic occupancy grids. However, so far, occupancy grids have been generated using depth information generated by distance sensor [9, 13–17], stereovision [5, 8] or deep learning [5]. Kim et al. [18, 19] build an occupancy map based on a dense back projection on the ground plane using multiple PoV, considering the ground plane to be at $Z = 0$. In this paper, we are focusing on generating semantic occupancy grids with the objective of increasing safety and facilitating the navigation in intersections and roundabouts. In order to limit the impact on the network performance of the cooperative system, we have decided to rely exclusively on bounding boxes given by on-the-shelf solutions such as YOLO [20] or the Mobileye solution used in [21].

The fusion of occupancy grids is generally based on the Bayesian framework as in [7, 22]. However, another approach formalized by Dempster [23] and completed by Shafer [24] called the Dempster-Shafer Theory (DST) is also used to perform occupancy grids fusion [5, 9]. Both frameworks, namely the Bayesian and the DST ones, will be investigated in this work in order to assess the best fusion approach to our context.

On the other hand, one of today's major challenges is to provide information to the decision toolset in charge of driving [25, 26]. This information used to be provided by the vehicle itself through its embedded sensors and the vehicle was therefore completely in charge of the perception task [27]. During the last decade, cooperative perception has increasingly been used, following the two main schemes of cooperation: the Vehicle-To-Vehicle (V2V) where vehicles share information with each other as in [28], and the Vehicle-To-Infrastructure (V2I) where the infrastructure shares scene perception information as in [13]. The cooperative perception offers, indeed, a solution to extend the fields of view beyond the limits of the embedded sensors and allows to reduce the occlusion effect which motivated us to undertake a cooperative perception-based approach.

2 System Architecture

In this section, we present our cooperative Vehicle-To-Everything (V2X) approach, collecting information from vehicles to build the map in order to broadcast the final generated semantic map to all agents in the scene. This approach allows the generation of semantic occupancy grids from sparse data (bounding boxes) gathered from cameras that are either placed on vehicles or on the infrastructure. The redistribution of the final map is out of the scope of this article.

2.1 Merging V2V and V2I Approaches

Today, cooperative perception projects are almost exclusively based on V2V or V2I paradigms [27]. In the V2I architecture, the infrastructure performs the perception task only with its sensors. This approach takes advantage of the elevated PoV offered by the infrastructure to reduce occlusions but neglects the in-the-scene PoV that can refine the results. Therefore, an approach using data from both the infrastructure and the connected vehicles navigating inside the scene is presented.

To make this system as versatile as possible, the infrastructure and vehicle PoV are considered as agents. All these agents send the bounding boxes' information along with the camera's one (sensor's pose and camera matrix) to the infrastructure. This data, encapsulated in a package, is then received by the RSU in order to perform the mapping of the scene and transmit the map afterward to the different users as presented in [2].

2.2 Road Side Unit Architecture

The RSU is the central element of our proposal. It processes, merges, and creates a semantic grid map from the data sent by the agents. Figure 1 shows the path of the data through the RSU and its different processing blocks up to the creation of the final map.

There are two sets of blocks. The former is made of the back projector, rasterizer, and BxA (Basic Probability/Belief Assignment) blocks, which are intended to perform the first processing on the data sent by each agent. In fact, for each agent, an instance of this first set is created and several instances may, therefore, be created in parallel. Since the output of this set has, not yet, benefited from the cooperative aspect, it is considered as local processing. This latter takes the form of a grid and will, therefore, be referred as a local

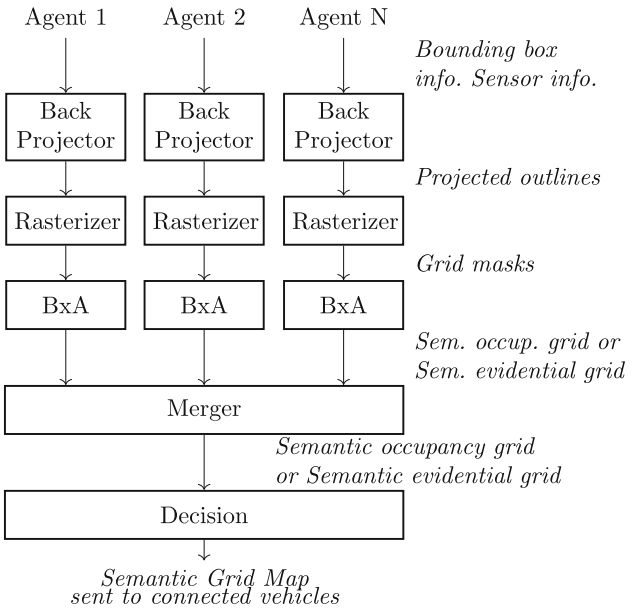


Fig. 1 RSU pipeline of the data received from the agent to create a semantic grid map. The illustration shows the example with 3 agents, and thus, 3 parallel processings before the merge of the grids

grid. The second set of blocks is intended to merge the local grids into a global semantic grid and thus, perform global treatments.

2.2.1 Local Processing Blocks

The local treatment consists of 3 blocks.

Back Projector This block uses the bounding box and sensor information to make an inverse projection of the bounding boxes onto the ground in the world frame.

Rasterizer It allows the creation of masks in the form of grids from the topological information of the previous block.

BxA This interchangeable block takes the format of Basic Probability Assigment (BPA) to convert the masks into a probabilistic occupancy grid or the format of Basic Belief Assigment (BBA) to convert the masks into an evidential occupancy grid.

2.2.2 Global Processing Blocks

The set performing the global treatment consists of two blocks depending on the type of the input grid.

Merger This block merges the local grids of each user using either a Bayesian or a DST based method.

Decision Finally, the global occupancy or evidential grid is converted into a semantic grid. This block must therefore make a decision about each cell belongs to which semantic class among a finite number of available semantic classes.

At the output of this set, a semantic grid map indicating where the objects are located is obtained. In the scope of this paper, only semantic classes of "pedestrians" and "vehicles" are considered. The other cells are considered as terrain, the default class. However, the number of classes can be extended to any number.

3 Local Grid Maps

In this section, we give details about the methods used in the three blocks of the local processing set.

3.1 Inverse Projection

To find the position of the users in the scene from the 2D bounding boxes, an inverse projection of the bounding boxes on the ground is performed. Indeed, the two bottom points of the 2D bounding box correspond approximately to the two closest points on the ground of the 3D bounding box, as shown in Fig. 2. The top two points of the bounding box, when they can be projected to the ground, give an upper limit to the span occupied by a user.

3.1.1 Plücker Coordinate System

In this section, we present a basic ray projection algorithm built upon Plücker coordinates as described in [29]. Plücker coordinates offer a powerful framework for representing lines and planes and finding their intersections in three-dimensional space. This ray projection algorithm allows us

Fig. 2 Bounding boxes for cars and pedestrians with their two lower points on the ground as given from our Dataset built from CARLA. The green frame represents the 2D bounding boxes extracted from the 3-Dimentional (3D) bounding boxes (in blue) given by the simulator

to perform inverse projections to retrieve the footprints of the detected vehicles on the ground plane.

Plane Let π a plane in homogeneous coordinates by Eq. 1.

$$\begin{aligned} \pi &= (\pi_1, \pi_2, \pi_3, \pi_4)^\intercal \\ \pi_1 X + \pi_2 Y + \pi_3 Z + \pi_4 &= 0 \end{aligned} \tag{1}$$

In other words, π_1, π_2, π_3 are the coordinates of the normal vector of the plane, and π_4 is the distance between the origin O and the plane π. Therefore, the vector π is built using the normal vector of the sought plane and its distance from the origin such that $\pi = [N|d]$ where $N \in \mathbb{R}^3$ is the normal vector of the plane and d is the distance between the origin O and the plane π, as defined in Eq. 2.

$$\begin{aligned} N &= (\pi_1, \pi_2, \pi_3)^\intercal \\ ||O\pi|| &= \pi_4 \end{aligned} \tag{2}$$

Line A ray, or line, can be defined by two points in homogeneous coordinates such that $A = [x_1, y_1, z_1, 1]^\intercal$ and $B = [x_2, y_2, z_2, 1]^\intercal$. The definition of the line from these two points is defined by Eq. 3.

$$L = AB^\intercal - BA^\intercal \tag{3}$$

Intersection To find the points in the 3D world that interest us, we look for the points of intersection between the ray L and the ground plane π. The coordinates of this point $P_{intersection}$ are obtained by Eq. 4 in non-normalized homogeneous coordinates.

$$P_{intersection} = L\pi \tag{4}$$

3.1.2 Inverse Projection

Now that we have a tool to perform the ray tracing task, we need to find the rays that pass through the corners of the bounding box and the center of the camera.

Pinhole Camera To make an inverse projection, we will have to see first the projection model of the objects in the 2D plane of the image. For that, the pinhole model as defined in [29] is considered. This model, defined in Eq. 5, allows to project a 3D point in the camera frame of coordinates $P_{cam} = (X, Y, Z)^\intercal$ on the image plane with coordinates $p_{img} = (u, v, w)^\intercal$ after normalization by the value of w.

$$p_{img} = K P_{cam} \tag{5}$$

where K is the camera matrix defined in Eq. 6 and constructed from f, the focal length, c_x and c_y the coordinates of the camera optical center on the image.

$$K = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \tag{6}$$

However, it should be noted that this model ignores the optical deformations that the lenses can bring. In this paper, these optical deformations are considered neglectable.

Inverse Pinhole Camera To go from a 2D point to a 3D point in the camera frame, the inverse principle is considered. Thus, for a point $p_{img} = (u, v, w)^\intercal$ on the plane, there exists a point $P_{cam} = (X, Y, Z)^\intercal$ in the camera frame, as given in Eq. 7.

$$P_{cam} = K^{-1} p_{img} \tag{7}$$

However, since we only have the coordinates u and v of the point in the image and the value of w is lost, the values of P_{cam} will depend on w. Therefore, instead of having a fixed point, a line defined by all values of w passing through the center of the camera and the real point in the world P_{real} is obtained. Therefore, a ray R_p is built from the point corresponding to the center of the camera, which will be named C_{world}, in the world frame, as well as a reprojected point of p_{img}, P_{cam}, with an arbitrary value of w, in the world frame and named P_{world}, after transformation by ${}^W T_C$, the transformation matrix from the camera frame to the world frame according to Eq. 8.

$$\begin{aligned} & R_{real} = (C_{world} P_{real}) \\ \forall w, \exists P_{cam},\ & P_{world} = {}^W T_C P_{cam} \in R_{real} \\ & \Rightarrow R_p = C_{world} P_{cam}{}^\intercal - P_{cam} C_{world}{}^\intercal \end{aligned} \tag{8}$$

Silhouette's Estimation We now have rays R created from the corners of the 2D bounding boxes and the center of the camera. The silhouettes are thus formed by these rays coming from the four corners of each of the bounding boxes and the ground plane π_{sol} according to Eq. 9.

$$P_{sol} = R\pi_{sol} \tag{9}$$

If a corner of a bounding box is above the horizon, then it will be projected to the infinity of the map. For this, the $< X, Y >$ coordinates of P_{world} for a w greater than the map size are taken.

However, we observe that the silhouettes projected on the ground are much larger than the span of the vehicle,

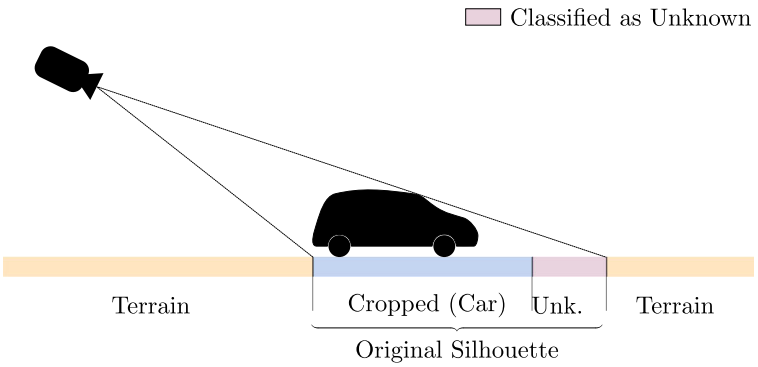
especially on the axis of the depth relative to the cameras. Therefore, this effect can be reduced by assigning a maximum length on the depth axis according to the class. In this case, a length of 6m for vehicles and 1m for pedestrians is chosen. The part of the original silhouette being trimmed will be considered hidden and therefore in an unknown state. This strategy is notably illustrated in Fig. 3.

3.2 Rasterization

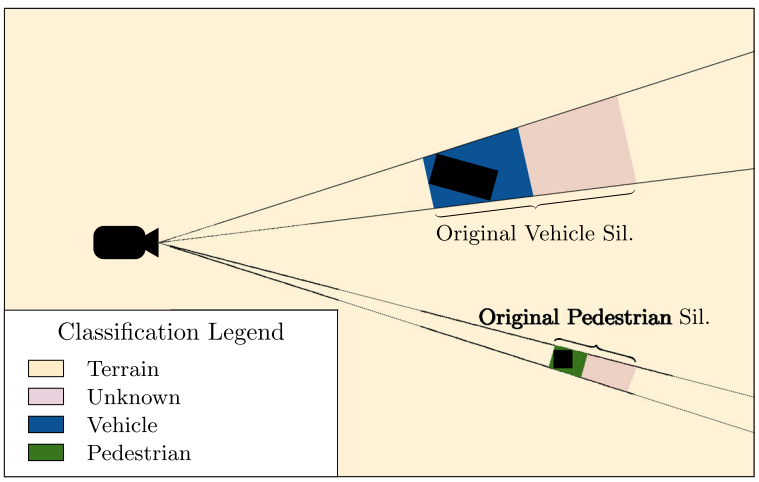
Since the silhouettes are in topological format, it will be difficult to merge them together. Therefore, let's convert this topological information into volumetric information. Among other things, either occupancy grids or obvious grids are used, depending on the desired fusion method. However, before obtaining an occupancy grid, it is necessary to rasterize the silhouettes. This step consists in defining for each cell if it belongs to a silhouette, to terrain or if it has not been observed.

First, all the cells are considered as unobserved. Then, the whole area covered by the camera is considered as terrain. To define this area, the principle explained in Section 3.1 is used with the 4 corners of the image. Finally, the cells belonging to a silhouette inherit its label (vehicle or pedestrian).

The map resulting from this operation thus takes the format of a grid where each cell contains a label (unknown, terrain, vehicle, or pedestrian). This grid can be denoted $M_{<x,y>}$ where $< x, y >$ are the cell coordinates. Since the latter has a similar structure to the images, the tools offered by the image processing library are used to perform the rasterization task. To plot the silhouettes on the occupancy grid, the function `fillPoly` of the OpenCV API [30] is used. This function is parametrized with the 8-connected lines mode, also called Moore's neighborhood to draw the polygons constituting the silhouettes. This mode takes into account the 8 cells bordering around a cell to draw a line, contrary to the 4-connected. In this mode, there is no antialiasing in order to have only one label per cell. Indeed, using a method per-



(a) Side View displaying the case where the originally projected silhouette is cropped to reduce its size to an acceptable dimension. The removed part of the silhouette is considered hidden by the object observed and thus classified as unknown.



(b) Top view displays the original silhouettes for a vehicle and a pedestrian and the cropped ones. In the case of a pedestrian, the dimensions to crop a silhouette are smaller than for the vehicles.

Fig. 3 The rays of the bounding boxes are projected onto the ground. If the silhouette is too large, it is reduced along its length. The areas resulting from the reduction are considered as unknown since they are occluded

forming antialiasing would mean applying probabilities to each label for each cell. However, depending on the merging mode, the map format differs.

3.3 Basic Assignment

The task of putting the label grids into a compatible format for merging belongs to the BxA block. If this block creates a semantic occupancy for the Bayesian-based merging method, it is then called BPA. However, if it uses a mass system used in DST to create an evidential grid, it is then called BBA.

3.3.1 Classes

In our work, we use three semantic classes, considered as possibilities, namely: pedestrian, vehicle, and terrain as given in Eq. 10,

$$\Omega = \{\mathcal{V}, \mathcal{P}, \mathcal{T}\} \tag{10}$$

where $\mathcal{V}$ is the vehicle class, $\mathcal{P}$ is the pedestrian class, and $\mathcal{T}$ is the terrain class. Ω represents the available universe of classes. In addition, there is an internal state used to define whether a cell has been observed or not. This will be treated differently depending on the merge mode, Bayesian or evidential.

3.3.2 Occupancy Grids

In the case of merging the different PoV by a Bayesian method, the previously generated grid is then transformed into a semantic occupancy grid.

Grid Definition This type of grid has already been defined, as in [12] where the authors propose an augmentation of the classical occupancy grid by appending the presumed class to the occupancy value. Nevertheless, this format is not suitable for grid fusion. Therefore, the format presented in [10] is chosen which, for each position, proposes $|\Omega|$ sub-cells, containing the probability of each class. We will note this map $\mathcal{B}_{<x,y,c>}$ with $< x, y >$ the coordinates of the cell, c the index of the sub-cell (one per class).

Basic Probability Assignment Function The probability value assignment is done based on the observed cell label and the detection confusion estimate. This task is here called the function BPA and can be defined according to Eq. 11.

$$BPA : M_{<x,y>} \rightarrow \mathcal{M}_{<x,y,z>},\ z \in \Omega \tag{11}$$

To perform this task, we use a lookup table that allows us to know the probability value of each class for each observed label. Table 1 shows the Look-Up Table (LUT) used for observations from vehicles. In this example, when a vehicle has been detected, we estimate the fact that it is really a vehicle at 85, that it is finally a pedestrian at 10 and that it is a land at 5. Table 2 shows the LUT used for observation from the infrastructure.

Table 1 LUT to assign probability values to each sub-cell based on the observed class of the original cell when observed from a vehicle. X stands for unobserved cases

Obs.	$\mathcal{V}$	$\mathcal{P}$	$\mathcal{T}$
X	0.33	0.33	0.33
$\mathcal{V}$	1.00	0.00	0.00
$\mathcal{P}$	0.00	1.00	0.00
$\mathcal{T}$	0.20	0.20	0.60

3.3.3 Evidential Grids

To perform a merge in the framework of the DST, it is necessary to create evidential maps.

Grid Definition The map takes a format very similar to the semantic occupancy grids presented in Section 3.3.2 but has more sub-cells than $|\Omega|$. In fact, in DST theory, the set of classes presented in Eg. 10 is augmented by considering the 2^{Ω} which is defined in Eq. 12. Thus, the cells are made of $|2^{\Omega}|$ sub-cells. This evidential grid format was notably used by Richter et al. in [5]. This map is noted $\mathcal{E}_{<x,y,c>}$ with $< x, y >$ the coordinates of the cell, c the index of the sub-cell (one per element of the power set).

$$\begin{aligned} 2^{\Omega} = \{\varnothing, \{\mathcal{V}\}, \{\mathcal{P}\}, \{\mathcal{T}\}, \\ \{\{\mathcal{V}\}, \{\mathcal{P}\}\}, \{\{\mathcal{V}\}, \{\mathcal{T}\}\}, \{\{\mathcal{P}\}, \{\mathcal{T}\}\}, \Omega\} \end{aligned} \tag{12}$$

The advantage of using a power set is that we can take into account states of doubt. For instance, in the case where a motorcycle is seen from the front, it could be classified as a pedestrian while seen from the side it would be more easily classified as a vehicle. It is thus possible to compute specifically the confusion factor between these two classes to assign a mass value to the set $\{\{\mathcal{V}\}, \{\mathcal{P}\}\}$.

Masses The mass values as presented in the previous paragraph, quantify evidence for each set of the power set. The mass function that attributes mass values to each set is defined in Eq. 13.

$$\begin{aligned} m : 2^{\Omega} &\rightarrow [0, 1] \\ m(\varnothing) &= 0 \end{aligned} \tag{13}$$

Table 2 LUT to assign probability values to each sub-cell based on the observed class of the original cell when observed from the infrastructure. X stands for unobserved cases

Obs.	$\mathcal{V}$	$\mathcal{P}$	$\mathcal{T}$
X	0.33	0.33	0.33
$\mathcal{V}$	1.00	0.00	0.00
$\mathcal{P}$	0.00	1.00	0.00
$\mathcal{T}$	0.00	0.00	1.00

Table 3 LUT to assign mass values to each sub-cell from the observed class of the original cell when observed from a vehicle

Obs.	$\varnothing$	$\{\mathcal{V}\}$	$\{\mathcal{P}\}$	$\{\mathcal{T}\}$	$\Omega\backslash\{T\}$	$\Omega\backslash\{P\}$	$\Omega\backslash\{V\}$	Ω
X	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
$\mathcal{V}$	0.00	0.30	0.00	0.00	0.10	0.10	0.00	0.50
$\mathcal{P}$	0.00	0.00	0.30	0.00	0.10	0.10	0.00	0.50
$\mathcal{T}$	0.00	0.10	0.10	0.30	0.00	0.00	0.00	0.50

X stands for unobserved cases

When associating values with masses, it is necessary to follow the property of Eq. 14.

$$\sum_{A\in 2^{\Omega}} m(A) = 1 \tag{14}$$

These are the masses that are stored in the sub-cells of the evidential grids.

Basic Belief Assignment Function In the same way, as for the function of BPA, the function of BBA allows to determine values for each of the masses of the power set and can be formalized in the form of Eq. 15:

$$BBA : M_{<x,y>} \rightarrow \mathcal{M}_{<x,y,z>},\ z \in 2^{\Omega} \tag{15}$$

Similarly to the occupancy grids, we also use a LUT similar to that of Table 3 to assign values to the masses of the power set depending on the observation of the cell when observed from a vehicle. We observe, however, that when a cell has not been observed, the mass of Ω is assigned to 1 to account for this state of unknown, unlike the BPA function. Table 4 shows the LUT used for observations from the infrastructure.

Today, the functions of BBA still form contributions since no method is yet agreed upon. Thus, we have defined the values of our LUT with a heuristic method using qualitative and quantitative studies. Another good indicator is the conflict value used in the Dempster fusion rule given in Eq. 21 which should be minimal. However, since these values are influenced by the performances of the agents' classifiers, but also by their pose, their number or by the layout of the terrain in the scene, it is necessary to frequently reevaluate the LUT's values.

Table 4 LUT to assign mass values to each sub-cell from the observed class of the original cell when observed from the infrastructure

Obs.	$\varnothing$	$\{\mathcal{V}\}$	$\{\mathcal{P}\}$	$\{\mathcal{T}\}$	$\Omega\backslash\{T\}$	$\Omega\backslash\{P\}$	$\Omega\backslash\{V\}$	Ω
X	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
$\mathcal{V}$	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.60
$\mathcal{P}$	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.60
$\mathcal{T}$	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.60

X stands for unobserved cases

4 Merging Methods

In this section, we come back to the merger block by detailing its functioning and the different approaches evaluated. In fact, we consider two main approaches: one based on Bayesian theory and the other on DST. Figure 4 illustrates these two approaches and the rules considered for merging the maps input to this block. For the Bayesian method, we used joint probabilities and thus a succession of multiplications, as well as the method based on the Sum Over the Log-Odds (SOLO). In the DST case, we considered two combination rules: the conjunctive rule and Dempster's rule. We firstly describe the method based on the Bayesian theory then the method based on the DST one.

4.1 Bayes-Based Merging

The first method implemented is the Bayesian fusion method, as proposed by the authors of [7, 22]. This method consists in using probability theory to estimate the probability that two images are similar.

Agents perform perception independently, *i.e.*, they do not take into account the observations of other agents to define the bounding boxes to be detected and they do not take into account past observations. Hence, we can make the assumption that there is no dependency between different observations of a given cell. The joint probability of a cell belonging to a given semantic class that is being observed by

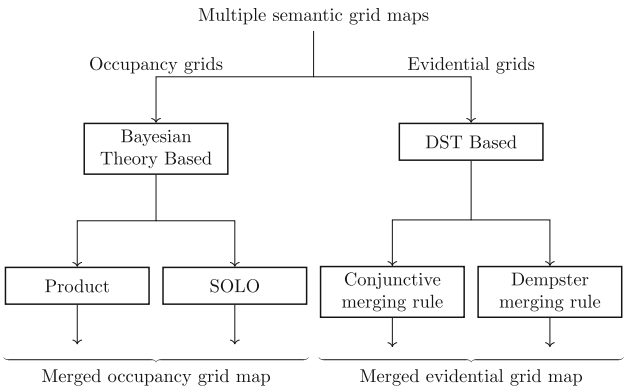


Fig. 4 Investigated merging approaches for occupancy and evidential grids. For both Bayesian theory and DST based approaches, two merging rules have been considered

two agents where agent 1 performs an observation denoted as o_1 and agent 2 as o_2 is expressed in Eq. 16.

$$P(o_1 \cap o_2) = P(o_1) \times P(o_2) \tag{16}$$

Since this operation is associative, for N agents, the probability associated with a sub-cell of the global map $\mathcal{M}_{\mathcal{B}<x,y,c>}$ can be computed from the maps issued from the agents $\mathcal{B}_{<x,y,c>i}$ according to Eq. 17, i being the index of the agent, and knowing that the cell contains the probability of the presence of its associated label.

$$\forall x \in [0,m],\ y \in [0,n], c \in \Omega$$
$$\mathcal{M}_{\mathcal{B}<x,y,c>} = \prod_{i}^{N} \mathcal{B}_{<x,y,c>i} \tag{17}$$

Therefore, for each subcell at a given $< x, y, c >$ coordinates, the observations can finally be merged by successive multiplications. Another method, notably used to reduce the approximation in floating point representation in the case of successive operations, is the use of the SOLO as given in [31]. Nevertheless, none of these methods handle observation conflicts.

4.2 Evidential Merging

A method based on DST as used in [5, 9] provides a better understanding of conflicting observations. Several combination rules are available.

4.2.1 Conjonctive's Combination Rule

The first combination rule, called the conjunctive combination rule, is defined by Eq. 18,

$$m_1(A) \textcircled{\cap} m_2(A) = \sum_{B \cap C = A \in 2^{\Omega}} m_1(B) m_2(C) \tag{18}$$

where m_1 and m_2 are mass functions defined over the universe Ω. Since the combination rule is associative, we can apply it to the $\mathcal{E}_{<x,y,c>}$ maps of each of the N agents to form a global evidential grid $\mathcal{M}_{\mathcal{E}<x,y,c>}$ according to Eq. 19:

$$\forall x \in [0,m],\ y \in [0,n], c \in \Omega$$
$$\mathcal{M}_{\mathcal{E}<x,y,c>} = \bigcap_{i=0}^{N} \mathcal{E}_{<x,y,c>i} \tag{19}$$

Following the association of the local grids, a global grid is obtained with the particularity of having $m(\varnothing) \neq 0$ in some cells. This value is generated by conflicts between the different agents observing the same cell. Several interpretations of the conflict are possible [32] such as the non-exhaustiveness of the discernment framework (lack of available classes), lack of reliability in the observations, or bad modeling of the perception capacities (BBA). Therefore it can be a good indication of the weaknesses of our modeling of the scene and of the perception that we will try to correct in order to reduce the conflict. Nevertheless, it is sometimes impossible to reduce this conflict and it will have to be managed either by coefficients of collapse in the BBA or in the phase of combination.

4.2.2 Dempster's Combination Rule

To handle the conflict in the combination phase, a normalization factor can be added to the conjunctive combination rule to form the Dempster combination rule. This is formalized in Eq. 20,

$$m_1(A) \oplus m_2(A) = \frac{1}{1-K} \sum_{B \cap C = A \neq \varnothing} m_1(B) m_2(C) \tag{20}$$

where K, defined in Eq. 21 gives the conflict value.

$$K = \sum_{B \cap C = \varnothing} m_1(B) m_2(C) \tag{21}$$

Thus, using Dempster's combination rule, the conflict is distributed among all masses but respects $m(\varnothing) = 0$, a property that must be respected in the closed world proposed by Shafer.

Moreover, this rule is also associative. Hence, it is possible to create a map from N observing agents providing local evidential grids $\mathcal{E}_{<x,y,c>}$ in order to obtain a global evidential one $\mathcal{M}_{\mathcal{E}<x,y,c>}$ according to Eq. 22.

$$\forall x \in [0,m],\ y \in [0,n], c \in \Omega$$
$$\mathcal{M}_{\mathcal{E}<x,y,c>} = \bigoplus_{i=0}^{N} \mathcal{E}_{<x,y,c>i} \tag{22}$$

At this point, either a semantic occupancy grid $\mathcal{M}_{\mathcal{B}<x,y,c>}$ or an evidential semantic grid $\mathcal{M}_{\mathcal{E}<x,y,c>}$ can be obtained. These maps contain the information for each class, but it is necessary to interpret them to obtain a semantic grid representing the scene.

5 Decision Methods

In this section, we discuss the decision block which, starting from a map in occupancy grid or semantic evidential format, generates a semantic grid. Figure 5 illustrates the method

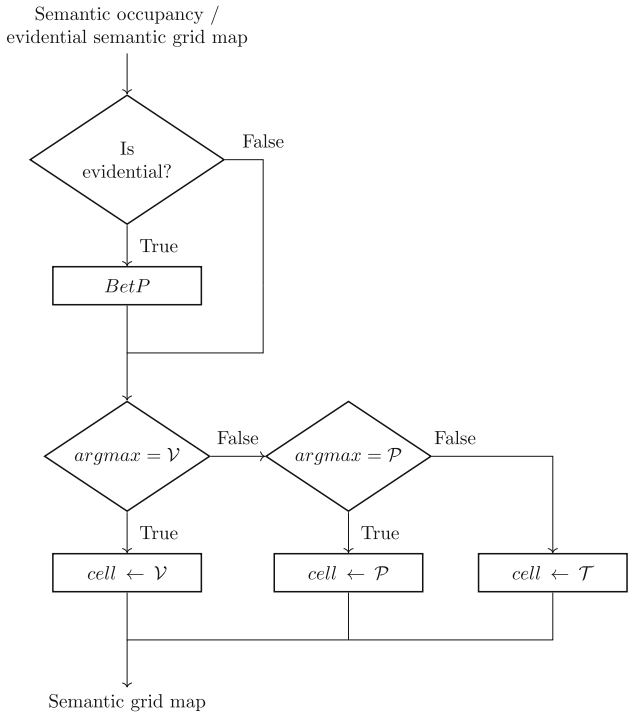


Fig. 5 Description of the transformation of semantic evidential or occupancy grid maps into a semantic grid. When the input map is an evidential map, it is transformed into a pignistic probability map. The semantic map then constructed is made up of cells stamped with the element with the highest probability

used to obtain the semantic map. In the following subsections, we formalize the transition from a semantic occupancy grid to a semantic grid. We then formalize the transformation of the semantic evidential grid into a semantic grid.

5.1 From Occupancy Grids To Semantic Grids

As defined in Section 3.3.2, the occupancy grid $\mathcal{M}_{\mathcal{B}<x,y,c>}$ consists of subcells containing the probability associated with each label. Thus, it is possible to transform the occupancy grid into the semantic grid $\mathcal{S}_{<x,y>}$ by selecting the label with the highest probability as formalized in Eq. 23.

$$\begin{aligned} &\forall x,\, y \in [0,m],[0,n] \\ &\mathcal{S}_{<x,y>} = \underset{c\in\Omega}{\text{argmax}}\, \mathcal{M}_{\mathcal{B}<x,y,c>} \end{aligned} \tag{23}$$

$\mathcal{S}_{<x,y>}$ thus consists, for each location cell $< x, y >$, of the label with the maximum estimated probability.

5.2 From Evidential Grids To Semantic Grids

It is possible to determine the pignistic probability noted $BetP$ of a label $A \in \Omega$ using Eq. 24.

$$BetP(A) = \sum_{\varnothing \neq B \subseteq \Omega} \frac{m(B)}{1-m(\varnothing)} \frac{|A \cap B|}{|B|},\ \forall A \subseteq \Omega \tag{24}$$

The advantage of calculating the pignistic probability resides in its consideration of the conflict estimation, defined in the Section 4.2.1, in the decision-making.

The method to define the map is based on the maximum pignistic probability among the elements of Ω, such as Eq. 25.

$$\begin{aligned} &\forall x,\, y \in [0,m],[0,n] \\ &\forall c \in 2^{\Omega},\ m(c) = \mathcal{M}_{\mathcal{E}<x,y,c>} \\ &\mathcal{S}_{BetP<x,y>} = \underset{C\in\Omega}{\text{argmax}}\ BetP(C) \end{aligned} \tag{25}$$

6 Results

In this section, we evaluate our approach. We first present the data used for the evaluation, then the metrics allowing a quantitative evaluation. Finally, we discuss the performance of our cooperative semantic map creation approach.

6.1 Datasets

In order to evaluate our algorithm, it is necessary to put it in situations which is possible via the use of datasets. In [5], the authors based their evaluation on the KITTI dataset [33]. Nevertheless, the KITTI dataset is not a cooperative dataset and, to the best of our knowledge, no cooperative dataset was available.

6.1.1 CARLA

Since cooperative datasets are difficult to realize due to the synchronization and the pose estimation of all the actors' requirements, we have chosen to realize a cooperative dataset from CARLA [34]. We noticed in parallel to our work that other teams have also realized cooperative datasets based on CARLA. This is notably the case of the authors of [35] who propose OPV2V, a cooperative dataset to test V2V approaches. Nevertheless, our approach also requires views from infrastructures.

(a) Traffic at an intersection featuring 4 infrastructure PoVs, 18 vehicles with embedded cameras and 20 pedestrians.

(b) Low density traffic at an intersection featuring 1 infrastructure PoV, 3 vehicles with embedded cameras and no pedestrian.

(c) Dense traffic at an intersection featuring 6 infrastructure PoVs, 30 vehicles with embedded cameras and 6 pedestrians.

Fig. 6 Capture of 3 situations generated with the CARLA simulator [34]

6.1.2 Our Dataset

To evaluate our approach, a dataset with PoV from vehicles and roadside infrastructure has been made. These PoV are instrumented identically with an RGB camera and a semantic camera. The position of the sensors and the position of the bounding boxes of the actors are also recorded for each simulation step. The actors are vehicles and pedestrians using the CARLA autopilot.

Since CARLA provides only the 3D bounding boxes of the object, we projected the 3D bounding boxes to the image [29]. Some of the 2D bounding boxes should not appear because occluded by other objects. We use the semantic segmented images associated with the RGB images to figure out the ratio of correct labels within a bounding box in order to define if the object is occluded and the bounding box is erased. A shortcoming of this solution is that objects occluded by a same-label object are not erased as visible in Fig. 2. Finally, an adjustable noise can be added to the retained bounding boxes.

To observe different behaviors, we generated several datasets with different traffic densities at a roundabout such as illustrated in Fig. 6b and in Fig. 6c. We also generated another dataset at a crossroads, illustrated in Fig. 6a. Our goal is to test the performance of our approach in several situations where there may be occlusions or confusion among agents. Our dataset can be augmented by enabling or disabling agents. By default, all vehicles are considered agents and provide a stream of images. It is, therefore, possible to ignore image streams to simulate vehicles that are not contributing.

6.2 Qualitative Study

To ensure that our system could generate coherent and usable maps, we conducted a qualitative study. We also used this qualitative assessment to roughly adjust the parameters used in the BBA and BPA. Figure 7 illustrates a visual comparison between the ground truth Fig. 7b, the map generated using the DST Fig. 7a, and the map generated using a Bayesian fusion-based approach Fig. 7c. The Bayesian theory-based method succeeds in placing all vehicles on the map, as does the DST based method. However, the method based on the DST seems to have fewer false positives. As for the pedestrian, they are mostly correctly placed on the map.

In order to compare our results, we tried to reconstruct a scene using a Multi-View Stereo approach as proposed in COLMAP [36] to find obstacles and build a map. However, despite the fact that we gave the true positions of the cameras in the a priori, more margin of error to the algorithm or pairs of initial images, we could not get any results. The reason is the lack of common features between the elements which underlines that a feature matching based approach is not adapted for perception systems featuring large baselines. We do not know any other method to create an occupancy grid from images without depth information from multiple POVs with freely available code. In contrast, our pragmatic and efficient approach is robust to large baselines.

6.3 Metrics

To provide a quantitative study, we used several metrics commonly found in the literature. Thus, we chose to use the Intersection over Union (IoU) as well as the F1-score to measure the performance on the size and the detection of objects. To measure the semantic performance, we used the Correct Ratio (CR).

6.3.1 Intersection over Union

The IoU is based on the number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). They are generated by comparing the cells of the ground truth map with the obtained semantic map. In this case, for a label $\omega \in \Omega$, if a cell of the semantic map obtained is equal to ω and is the same on the cell of the same position on the ground truth map, then this cell is regarded as a TP. If a cell of the obtained semantic map is equal to ω but it is not equal to ω on the cell of the same position on the ground truth map, then this cell is considered as an FP. If a cell of the obtained semantic map is not equal to ω but is equal to ω on the cell of the same position on the ground truth map, then this cell is considered as an FN. Finally, if a cell of the obtained semantic map is not equal to ω and it is not equal to ω on the cell of the same position on the ground truth map, then this cell is considered as an TN. Thus, the IoU for a chosen ω label is given by Eq. 26.

$$IoU_\omega = \frac{TP_\omega}{TP_\omega + FP_\omega + FN_\omega} \tag{26}$$

To estimate the overall performance, it is possible to calculate the average between all labels, as given in Eq. 27.

$$mIoU = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} IoU_\omega \tag{27}$$

However, we have noted a limitation of the average score. When the detection gives a failure rate of 100% and, therefore, the whole map is considered as terrain, the default label, then the average IoU is about 30%. Another solution is to transform the semantic grids into an occupancy grid and to compute the IoU on the occupancy rather than on the labels.

6.3.2 F1-Score

The F1-Score is very similar to the IoU since it is also based on the number of TP, TN, FP and FN. It can be calculated as shown in Eq. 28.

$$F1_\omega = \frac{TP_\omega}{TP_\omega + \frac{FP_\omega + FN_\omega}{2}} \tag{28}$$

In the same way, as for the average F1-score, it is possible to obtain an overall value by calculating the average F1-score, as in Eq. 29. It should be noted that the average F1-score shares the same shortcomings as the average IoU.

$$mF1 = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} F1_\omega \tag{29}$$

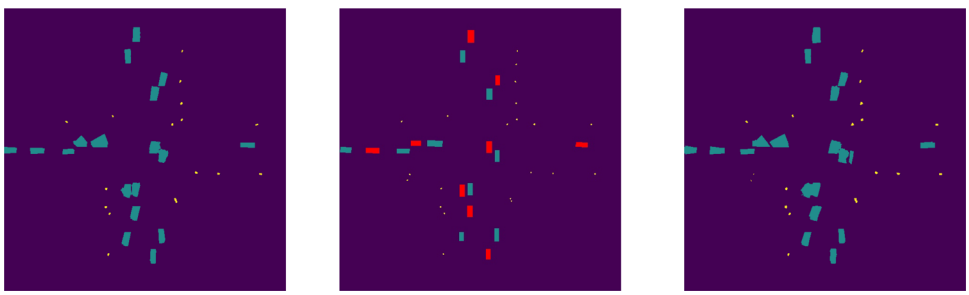


Fig. 7 Comparison of the ground truth maps and the semantic map generated by our solution. In purple: ground cells, in yellow: pedestrians, and in turquoise and red vehicles. Red vehicles show the connected vehicle distribution of 50% among the fleet

(a) Semantic map generated by our approach using DST merging.

(b) Ground Truth map.

(c) Semantic map generated by our approach using Bayes-based merging.

Table 5 Detail of the IoU, F1-Score and CR (in %) for an infrastructure hosting 6 PoV, 30 vehicles (all contributing) and 6 pedestrians

Fusion	Class	IoU	F1-Score	CR
Ours (Bayes) *Product*				
	$\mathcal{V}$	26.21	36.46	96.50
	$\mathcal{P}$	22.33	41.52	99.95
	$\mathcal{T}$	96.40	98.17	96.45
	Mean	48.31	58.71	N/A
Ours (Bayes) *SOLO*				
	$\mathcal{V}$	26.73	42.16	96.81
	$\mathcal{P}$	22.79	37.06	99.95
	$\mathcal{T}$	96.72	98.33	96.76
	Mean	48.75	59.18	N/A
Ours (DST)	$\mathcal{V}$	50.55	67.07	98.87
	$\mathcal{P}$	28.03	43.49	99.98
	$\mathcal{T}$	98.84	99.41	98.85
	Mean	59.14	69.99	N/A

6.3.3 Correct Ratio

In order to measure the performance on assigning correct labels to cells, we used the CR which we calculated as shown in Eq. 30.

$$CR_\omega = \frac{TP_\omega + TN_\omega}{TP_\omega + TN_\omega + FN_\omega + FP_\omega} \tag{30}$$

Usually, the CR is calculated by comparing corresponding label cells on the ground truth map and the final map divided by the total number of cells in the map. However, the cells having a correct label are constituted by the sum of TP and TN. The limitation of this metric in our use case is that the majority of the cells are considered as terrain in the map to be evaluated and in the ground truth map. Therefore, the results are always very high and it is difficult to distinguish the variations.

6.4 Quantitative Study

In this section, we observe our approach in terms of several parameters using the metrics designated above. All the results presented in the remainder of this section have been obtained on a grid map of 120m with square cells with a size of 0.2m size, centered at the barycenter of the infrastructure PoV positions.

6.4.1 Bayes-based Method vs. DST-based Method

Since we have tried two approaches, one based on Bayesian theory and the other based on the DST, we want to observe the differences in performance between the two approaches. Similarly with Fig. 7 displaying the maps generated by the two methods, Table 5 aims precisely at showing the performance differences between the two approaches for each of the metrics stated earlier which can be used as a reference point for the remainder of this article. Figure 8 shows the difference between the two approaches on the same dataset, based on a sample of 300 frames.

The results highlighted in Table 5 express an average improvement of 22.42% on the average IoU and 19.21% on the mean F1-Score in favor of the DST based approach. Vehicles benefit the most from this approach with a gain on the IoU of 92.87%. Pedestrians also benefit from a better representation on the map thanks to the approach based on the DST. However, we notice that the pedestrian IoU is low compared to the other classes. This is due to the fact that the areas of the cells are significant compared to the areas occupied by pedestrians. Thus, the number of cells occupied by pedestrians is low and artificially increases the impact of errors in the metrics. Conversely, for the terrain class which occupies the majority of the cells of the map and on which the impact of errors is particularly low in the metrics. Table 5 also displays the results given for the Bayesian approach using

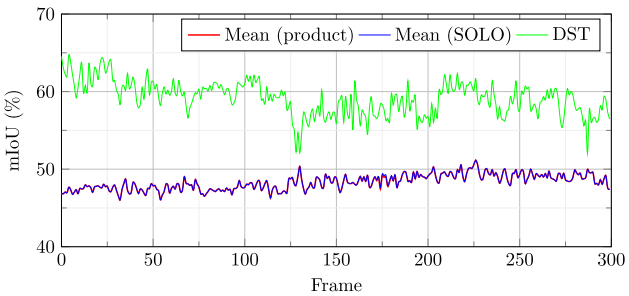


Fig. 8 Evolution of the mean IoU through 300 frames for our three fusion methods

Table 6 Evolution of the mIoU (in %) for the dense traffic scene (roundabout) of our dataset, varying the proportion of agents in the users fleet

Infrastructure: N# PoV	Connected Vehicles	Ours (Bayes) mIoU (%)	Ours (DST) mIoU (%)	Gain (%)
Single-view				
0 (0%)	1 (3%)	53.22	53.28	0.11
2 (33%)	0 (0%)	54.43	55.69	2.31
	1 (3%)	**54.85**	56.00	2.10
	8 (27%)	51.97	56.77	9.24
	15 (50%)	50.54	**57.64**	14.05
	23 (77%)	50.30	57.58	14.47
	30 (100%)	49.87	57.29	**14.88**
6 (100%)	0 (0%)	50.67	58.41	15.28
	1 (3%)	**51.24**	58.53	14.23
	8 (27%)	49.91	59.23	18.67
	15 (50%)	48.74	**60.12**	**23.35**
	23 (77%)	48.58	59.87	23.24
Full dataset (6 Infra. PoV + 30 CVs)	48.31	59.14	22.42	

Bold-faced numbers highlight the highest value and, thus, the configuration that brought the best result for the given measure

the SOLO [31]. However, compared to the results obtained with products, the improvement is not significant. Figure 8 reinforces these conclusions, highlighting a significant difference between the performance of the DST-based approach and that of the Bayesian-based approach. It also shows the similar output of the product and SOLO methods Thus, in the remainder of this article, we use the product methods which is less compute-intensive.

6.4.2 Connected Vehicles Ratio Evolution in a Scene

Now that we have seen the performance between the two approaches of our solution, we can test, on the same scene of our dataset, to vary the proportion of CV and PoV of the infrastructure. We, therefore, performed several subscenarios. The first one consists of a single vehicle observing the scene, as an instrumented vehicle alone. The second scenario consists of infrastructure alone in the manner of projects like [13]. A third scenario is to have the infrastructure with only 1 CV corresponding to the approach of MEC-View[3]. Finally, other scenarios are created by changing the proportion of CV up to the all connected. Figure 7b show the distribution of connected and unconnected vehicle in a scene for a proportion of 50% of connected vehicles.

Table 6 shows that the approach based on the Bayesian theory maintains a IoU of 50% and seems to suffer from the multiplication of the points of view whereas the approach based on the DST benefits more from the multiplication of the points of views. Indeed, in the scene of dense traffic in a roundabout, occlusions are frequent and can produce conflicting observations between the agents. However, the approach based on the DST manages the conflicting observations and thus shows its advantage in such scenarios, unlike the approach based on the Bayesian theory. Thus, as the number of PoV increases, the gap between the DST approach and the Bayesian approach widens, up to a maximum of 23.35% of mean IoU (mIoU) gain.

We also observe that with an infrastructure reduced to the strict minimum and a fleet with a proportion of about 50% of CV, it is possible to generate a map with good results. This observation is, therefore, encouraging in the transition that we will see until we have 100% of CV instrumented on the roads.

6.4.3 Traffic Density Evolution

Finally, another important variable at intersections is traffic density. Indeed, the denser the traffic is, the more the phenomenon of occlusions is accentuated and the more difficult the scene is to understand and map. We have three scenarios with several varying the number of agents at the same roundabout as shown in Table 7.

As in Tables 6 and 7 points out that the more observers there are, the larger the gap between the DST based approach and the Bayesian theory-based method. However, we also observe that even the DST approach is affected by the complexity of the scene due to occlusions and conflicting observations.

[3] http://www.mec-view.de/

Table 7 Comparison of the IoU varying the number of vehicles in the roundabout (in %)

Number of agents		4	12	36
Ours (Bayes)	$\mathcal{V}$	36.82	32.10	26.21
	$\mathcal{P}$	N/A	26.70	22.33
	$\mathcal{T}$	99.39	98.98	96.40
	Mean	45.40	52.59	48.31
Ours(DST)	$\mathcal{V}$	42.21	53.99	50.55
	$\mathcal{P}$	N/A	32.57	28.03
	$\mathcal{T}$	99.55	99.59	98.84
	Mean	47.25	**62.05**	59.14
Gain (%)	Mean	4.07	18.00	**22.42**

Bold-faced numbers highlight the highest mean IoU as well as the highest gain

Nonetheless, we observe that the values of mIoU are fair and that our solution provides usable maps regardless of the traffic density in the scene.

7 Conclusion

In this paper, we presented a new method to generate semantic grids from sparse and light information coming from both vehicle's embedded sensors and roadside infrastructure sensors. This approach is designed to be highly cooperative and exploits the in-scene PoV of the vehicles to refine the generated map.

Our pragmatic and efficient approach succeeded to generate maps regardless of the appearance of the objects from the multiple PoV and overtook other state-of-the-art tools such as COLMAP [36] which were unable to bring results due to the limitations of its algorithms based on depth and 3D reconstruction.

The presented method is based on two approaches: one based on Bayesian theory and the other based on the DST. The performances of our approach have been tested on a dataset composed of several scenes, generated with CARLA [34]. For a common road traffic scene, our method proposed a mIoU of 48.38% for the Bayesian-based method and 59.14% for the DST-based method, whereas in the same scenario, COLMAP failed to deliver any results demonstrating the value of our method for understanding a scene from multiple PoV and sparse information. Since a transition between our current world and a world where all vehicles are connected is inevitable, we simulated this transition by varying the ratio number of CVs in the scene. We observed an mIoU of 60.12% with the DST-based method when the scene is covered by 6 roadside PoV and 50% of connected vehicles. This represents a gain of 12.84% compared with the case where only one vehicle observes the scene with its on-board sensors. This gain of mIoU corresponds to the previously occluded cells where previously hidden road users are now observed by other PoV. This brings possible contributions of our method to road safety, by enabling the anticipation of other road users not yet visible to a driver. Finally, we tested the robustness of our approach in three scenarios with different traffic densities. We obtained a maximum mIoU of 62% for just 12 agents in the round-about; when the number of agents is lower, not all occluded areas are covered. As the number of agents increases, we observe a plateau phenomenon and a slight drop in mIoU due to conflicting observations, which explains the difference in performance between the Bayesian-based method and the DST-based one, which is more resilient when there are conflicting observations.

In future work, we will seek to develop a method to optimize the values of the LUT used in the BPA and the BBA as well as a method to measure the effectiveness of the LUT to trigger an automated recalibration. We will also run a study of the impact of the position and synchronization noise between the agents. In addition, a method for a better estimation of the agents pose using only bounding boxes without the depth information can be explored [37]. The temporal integration of the observation can also be explored and might bring finer results. Finally, the integration of other local semantic map building methods such as [38] to build the masks can be evaluated in order to determine if there is a significant impact on the performances compared to the network impact.

Acknowledgements This work is supported by ESIGELEC, Rouen, France through an internal Ph.D. grant.

Author Contributions All authors contributed to the study's methodology, conception and design. Formal analysis: Caillot Antoine and Ouerghi Safa; Software and experimental validation: Caillot Antoine, Ouerghi Safa and Dupuis Yohan; Resources: Caillot Antoine; Writing: The first draft of the manuscript was written by Caillot Antoine and all authors commented on previous versions of the manuscript, all authors read and approved the final manuscript; Supervision and administration: Ouerghi Safa, Dupuis Yohan, Pascal Vasseur and Boutteau Rémi.

Funding This work is supported by ESIGELEC, Rouen, France through an internal Ph.D. grant.

Availability of data and materials Basic datasets are available on the following address https://zenodo.org/record/7637904. More are to come.

Code availability The code is available on GitHub at the following address, https://github.com/caillotantoine/carla-V2X-dataset-generator, for the dataset generation. And, at the following address https://github.com/caillotantoine/Coop-Evidential-Semantic-Grid, for the multi agent grid generation.

Declarations

Conflicts of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethics approval Not applicable.

Consent to participate Informed consent was obtained from all participants.

Consent for publication Participant consented to the submission of this article.

References

1. Datondji, S.R.E., Dupuis, Y., Subirats, P., Vasseur, P.: A survey of vision-based traffic monitoring of road intersections. IEEE Trans. Intell. Transp. Syst. **17**(10), 2681–2698 (2016)
2. Caillot, A., Ouerghi, S., Vasseur, P., Dupuis, Y., Boutteau, R.: Multi-agent cooperative camera-based evidential occupancy grid generation. Conference on Intelligent Transportation Systems. (2022)
3. Elfes, A.: Using occupancy grids for mobile robot perception and navigation. Computer **22**(6), 46–57 (1989)
4. Thrun, S.: Learning occupancy grid maps with forward sensor models. Auton. Robot. **15**(2), 111–127 (2003)
5. Richter, S., Wang, Y., Beck, J., Wirges, S., Stiller, C.: Semantic evidential grid mapping using monocular and stereo cameras. Sensors. **21**(10) (2021). https://doi.org/10.3390/s21103380
6. Moravec, H., Elfes, A.: High resolution maps from wide angle sonar. In: Proceedings. 1985 IEEE International Conference on Robotics and Automation, vol. 2, pp. 116–121. IEEE (1985)
7. Birk, A., Carpin, S.: Merging occupancy grid maps from multiple robots. Proc. IEEE **94**(7), 1384–1397 (2006)
8. Jennings, C., Murray, D., Little, J.J.: Cooperative robot localization with vision-based mapping. In: Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C), vol. 4. pp. 2659–2665 (1999). IEEE. https://www.cs.ubc.ca/labs/lci/papers/docs1998/little-icra98.pdf
9. Camarda, F., Davoine, F., Cherfaoui, V.: Fusion of evidential occupancy grids for cooperative perception. In: 2018 13th Annual Conference on System of Systems Engineering (SoSE). pp. 284–290 (2018). https://doi.org/10.1109/SYSOSE.2018.8428723
10. Erkent, O., Wolf, C., Laugier, C.: Semantic grid estimation with occupancy grids and semantic segmentation networks. In: 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV). pp. 1051–1056 (2018). https://doi.org/10.1109/ICARCV.2018.8581180, https://hal.inria.fr/hal-01933939/document
11. Richter, S., Bieder, F., Wirges, S., Stiller, C.: Mapping lidar and camera measurements in a dual top-view grid representation tailored for automated vehicles. (2022). arXiv:2204.07887
12. Lu, C., Molengraft, M.J.G., Dubbelman, G.: Monocular semantic occupancy grid mapping with convolutional variational encoder-decoder networks. IEEE Robot. Autom. Lett. **4**(2), 445–452 (2019)
13. Annkathrin Krämmer*, Schöller*, D.G.C. , Knoll, A.: Providentia - a large scale sensing system for the assistance of autonomous vehicles. In: Robotics: Science and Systems (RSS), Workshop on Scene and Situation Understanding for Autonomous Driving. (2019). https://sites.google.com/view/uad2019/accepted-posters
14. Li, Z., Yu, T., Fukatsu, R., Tran, G.K., Sakaguchi, K.: Proof-of-concept of a sdn based mmwave v2x network for safe automated driving. In: IEEE Global Communications Conference (GLOBECOM). pp. 1–6. IEEE (2019)
15. Bosch: Conduite automatisée: comment les voitures et les infrastructures communiquent en milieu urbain. Website. Accessed 2024-03-22 (2020). https://www.bosch.fr/actualites/2020/conduite%2Dautomatisee%2Dcomment%2Dles%2Dvoitures%2Det%2Dles%2Dinfrastructures%2Dcommuniquent%2Den%2Dmilieu%2Durbain/
16. Gabb, M., Digel, H., Müller, T., Henn, R.-W.: Infrastructure-supported perception and track-level fusion using edge computing. In: IEEE Intelligent Vehicles Symposium (IV). pp. 1739–1745. (2019). https://doi.org/10.1109/IVS.2019.8813886
17. Lv, B., Xu, H., Wu, J., Tian, Y., Zhang, Y., Zheng, Y., Yuan, C., Tian, S.: Lidar-enhanced connected infrastructures sensing and broadcasting high-resolution traffic information serving smart cities. IEEE Access. **7**, 79895–79907 (2019)
18. Kim, S.-W., Chong, Z.J., Qin, B., Shen, X., Cheng, Z., Liu, W., Ang, M.H.: Cooperative perception for autonomous vehicle control on the road: Motivation and experimental results. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 5059–5066. IEEE (2013)
19. Kim, S.-W., Qin, B., Chong, Z.J., Shen, X., Liu, W., Ang, M.H., Frazzoli, E., Rus, D.: Multivehicle cooperative driving using cooperative perception: Design and experimental validation. IEEE Trans. Intell. Transp. Syst. **16**(2), 663–680 (2014)
20. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. (2018). arXiv:1804.02767
21. Baek, M., Jeong, D., Choi, D., Lee, S.: Vehicle Trajectory Prediction and Collision Warning via Fusion of Multisensors and Wireless Vehicular Communications. Sensors. **20**(1), 288 (2020). https://doi.org/10.3390/s20010288. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute. Accessed 2022-02-14
22. Franco, J.-S., Boyer, E.: Fusion of multiview silhouette cues using a space occupancy grid. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, vol. 2. pp. 1747–17532. (2005). https://doi.org/10.1109/ICCV.2005.105 ISSN: 2380-7504
23. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. J. Roy. Stat. Soc.: Ser. B (Methodol.) **39**(1), 1–22 (1977) https://doi.org/10.1111/j.2517-6161.1977.tb01600.x, https://www.rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1977.tb01600.x
24. Shafer, G.: Dempster-shafer theory. Encyclop. Artif. Intell. **1**, 330–331 (1992)
25. Xu, P., Dherbomez, G., Héry, E., Abidli, A., Bonnifait, P.: System architecture of a driverless electric car in the grand cooperative driving challenge. IEEE Intell. Transp. Syst. Mag. **10**(1), 47–59 (2018)
26. Englund, C., Chen, L., Ploeg, J., Semsar-Kazerooni, E., Voronov, A., Bengtsson, H.H., Didoff, J.: The grand cooperative driving challenge 2016: boosting the introduction of cooperative automated vehicles. IEEE Wirel. Commun. **23**(4), 146–152 (2016). https://doi.org/10.1109/MWC.2016.7553038
27. Caillot, A., Ouerghi, S., Vasseur, P., Boutteau, R., Dupuis, Y.: Survey on cooperative perception in an automotive context. IEEE Trans. Intell. Transpo. Syst. 1–20 (2022)

28. Kianfar, R., Augusto, B., Ebadighajari, A., Hakeem, U., Nilsson, J., Raza, A., Tabar, R.S., Irukulapati, N.V., Englund, C., Falcone, P., et al.: Design and experimental validation of a cooperative driving system in the grand cooperative driving challenge. IEEE Trans. Intell. Transp. Syst. **13**(3), 994–1007 (2012)
29. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision 2nd ed., 4th print. New York: Cambridge University Press (2006)
30. Bradski, G., Kaehler, A.: Opencv. Dr. Dobb's journal of software tools. **3**, 2 (2000)
31. Stachniss, C.: Exploration and mapping with mobile robots. PhD thesis, Citeseer (2006)
32. Lefèvre, E.: Fonctions de croyance: de la théorie à la pratique. PhD thesis (2012)
33. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. Int. J Robotics Res. (2013)
34. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: Proceedings of the 1st Annual Conference on Robot Learning. pp. 1–16. (2017)
35. Xu, H. Runsheng anpignistiqued Xiang, Xia, X., Han, X., Liu, J., Ma, J.: Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. (2021) arXiv:2109.07644
36. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.-M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV). (2016)
37. Mauri, A., Khemmar, R., Decoux, B., Haddad, M., Boutteau, R.: Real-time 3d multi-object detection and localization based on deep learning for road and railway smart mobility. J Imaging. **7**(8)(2021). https://doi.org/10.3390/jimaging7080145
38. Gan, L., Jadidi, M.G., Parkison, S.A., Eustice, R.M.: Sparse bayesian inference for dense semantic mapping. (2017) arXiv:1709.07973

Antoine Caillot received a Master of Engineering in Embedded Systems for Autonomous Vehicles from ESIGELEC (France, 2019). In 2022, he received his PhD degree from the University of Rouen Normandy (France) for works related to Robotics (Cooperative perception, Dempster-Shafer Theory). Since then, he is a postdoctoral fellow at the CNRS-AIST Joint Robotics Laboratory (JRL) in Tsukuba (Japan). His research interests include computer vision, Simultaneous Localization and Mapping (SLAM), cooperative perception and robotics.

Safa Ouerghi received her Master degree from SUP'COM (Higher School of Communication of Tunis) in Information and communication Technology in 2012. She received her PHD from SUP'COM in collaboration with ESIGELEC in 2018 for her work related to vision-based vehicle localization. From 2020, she is an Associate Professor at the ESIGELEC engineering school and a researcher in the IRSEEM research institute. Her research interests are perception, localization and computer vision dedicated to robotics and autonomous vehicles.

Yohan Dupuis received the MSc in Electrical Engineering from Union Graduate College, NY and a MEng from ESIGELEC, France, in 2009. In 2012, he earned a PhD in Computer Science from Université de Rouen. He is now Research Director at CESI LINEACT. His research interests focus on perception for ground vehicle-infrastructure interaction understanding.

Pascal Vasseur is full professor at the Université de Picardie Jules Verne (France) and is a member of the MIS laboratory. His research interests are computer vision and image processing and their applications to intelligent transportation, mobile and aerial robots.

Rémi Boutteau received his engineering degree from the IMT Lille Douai and his MSc degree in Computer Science from the University of Lille in 2006. In 2010, he received his PhD degree from the University of Rouen Normandy for works related to Computer Vision (catadioptric sensors, 3D reconstruction, Structure-from-Motion). From 2009 to 2020, he was an Associate Professor at the ESIGELEC engineering school and a researcher in the IRSEEM research institute. Since 2020, he is a Full Professor at University of Rouen Normandy within the STI team (Intelligent Transportation System) at the LITIS Lab (IT Laboratory, Information Processing and Systems). His research interests are perception, localization and computer vision dedicated to autonomous vehicles.

Authors and Affiliations

Antoine Caillot[1] · Safa Ouerghi[1] · Yohan Dupuis[2] · Pascal Vasseur[3] · Rémi Boutteau[4]

Safa Ouerghi
ouerghi@esigelec.fr

Yohan Dupuis
ydupuis@cesi.fr

Pascal Vasseur
pascal.vasseur@u-picardie.fr

Rémi Boutteau
remi.boutteau@univ-rouen.fr

[1] Normandie Univ, UNIROUEN, ESIGELEC, IRSEEM, Rouen 76000, Normandie, France

[2] CESI LINEACT, Paris La Défense 92060, Ile de France, France

[3] Modélisation Information et Systèmes (MIS), Université de Picardie Jules Verne, Amiens 80000, Hauts de France, France

[4] Univ Rouen Normandie, INSA Rouen Normandie, Université Le Havre Normandie, Normandie Univ, LITIS UR 4108, Rouen F-76000, Normandie, France