Intelligent vehicles semantic segmentation using evidential deep learning

Dănuţ-Vasile Giurgi[1], Mihreteab Negash Geletu[2], Thomas Josso-Laurain[1], Maxime Devanne[1], Jean-Philippe Lauffenburger[1], Mengesha Mamo Wogari[2]

[1] IRIMAS-UR7499 – Universite de Haute-Alsace, Mulhouse, France
[2] Addis Ababa University, Addis Ababa, Ethiopia
Email: vasile.giurgi@uha.fr

Abstract. Autonomous cars encounter momentous challenges in the perception tasks. The driving surrounding areas are more and more congested and the weather conditions differ significantly. Sensors-wise the capacities have increased, leading to an increasing interest in big data management such as artificial intelligence. Currently, neural networks have proved their efficiency, but restraints in complex situations are still present. In this work, a cross-fusion technique that combines lidar and camera data using an encoder-decoder-based model is proposed. The multi-modal architecture fuses different sources of information to circumvent encountered limitations. The considered perception task is semantic segmentation of the different obstacles that may be encountered. The decision-making part of the architecture is extended with the evidence theory, introducing belief functions that contribute in handling uncertainties. Thus, the evidential formulation is versatile and yields more precise predictions and a better understanding of the vacuous data. The dataset used in this work employs the KITTI dataset for semantic segmentation. The results show the interest of integrating evidential theory into neural networks fusing information from two heterogeneous sensors.

Keywords: cross-fusion, evidential deep-learning, perception, uncertainty.

1 Introduction

One important task in autonomous driving is to perceive the environment to achieve semantic analysis. This allows sending accurate information for the path planning and the control part to be applied. For the vision of the car, perception adopts substantially the deep learning-based approaches and the multi-sensor information to achieve strong capabilities. A big part of the actual self-driving car features rely on neural networks, especially deep learning, as their standard approach for perception tasks.

The popularity of this field soared when a deep convolutional neural network architecture named AlexNet outperformed other methods in a visual recognition challenge called ImageNet [1]. As a result, deep learning is now widely applied to various perception tasks in self-driving cars, including classification, object detection, semantic segmentation, and more [2, 3].

In this algorithm, the output of the prediction frequently depends on the sigmoid function for binary predictions and the softmax function for multi-class predictions. These outputs represent probabilities assigned to a group of exclusive prediction categories. Nonetheless, probabilities do not always consistently manage uncertainty effectively, as they can struggle to distinguish between the absence of information and conflicting information. Withal, performances are often affected, and probabilistic models can struggle to predict uncertainties and conflicting information. Hence for uncertainties, missing information, and imprecision, there are other effective techniques that exist such as the theory of evidence [4].

This theory, known as the Belief Functions (BF) theory or Dempster-Shafer theory (DS/DST), was initially proposed by Dempster and Shafer [4]. DS theory serves to represent belief elements for uncertain models and emphasizes several key features such as generality, operationality, or scalability [5]. In autonomous driving tasks, such as obstacle avoidance, belief functions have demonstrated their ability to provide accurate results. For instance, they improve the performance of LiDAR sensor occupancy grid maps by representing conflicts in a more meaningful manner [6].

In detecting pedestrians applications, evidential combination rules have consistently outperformed some Bayesian approaches [7]. Moreover, in multi-modal perception, evidential theory sets itself through the handling of missing information, imprecision, and ignorance.

In a multimodal approach, Xu et al. [8] integrated KITTI semantic segmentation images from various sensors, including cameras and different LiDAR layers. Their approach enables the expansion of object classes or the integration of additional sensors, resulting in improved performance and a better understanding of the drivable area.

In the context of self-navigation tasks relying on neural networks, as demonstrated in [9], a deep learning architecture based on Multi-Layer Perceptron (MLP) has been devised for the categorization of arbitrary LiDAR entities to enhance perception. This architecture departs from the probabilistic approach and adopts an evidential inference method, drawing inspiration from Denœux's generalized logistic classifier [10].

The application of evidence theory is employed to represent elements of evidence (beliefs) in the context of uncertain predictions by models. Consequently, the adoption of an evidential formulation shows promising outcomes (road segmentation and multi-object detection) in the realm of self-navigation embedded systems, which are the primary focus of this work.

The objective of this work is to develop an evidential deep learning model that integrates information from various sensors (camera and lidar) to enable autonomous driving capabilities. The work aims to offer decision-making rules that can explicitly produce cautious judgments to handle conflicts, whether originating from a single source of information or multiple sources (data fusion).

A cross-fusion model is augmented with evidence theory to achieve semantic segmentation performances within the KITTI dataset[11]. Thus, the workflow starts with a probabilistic approach, where the predictions are road, vehicles, and background and it is thereafter extended to an evidential formulation where the predictions are enlarged with a new class named ignorance (introduced by the the evidence theory), despite being trained on the road, vehicle, and background labels.

Through the leverage of theory-based belief functions, the model improves the decision-making part and allows representing uncertain prediction, which constitutes the primary contribution of this work.

The paper's organization continues through the subsequent sections: Background (covering the belief functions theory basics) Proposed method (illustrating some implementation details about the evidential formulation), Semantic segmentation results (encompassing data pre-processing and presenting results), and Conclusion.

2 Background

Evidence theory is a formal framework for making decisions in the presence of uncertainty. One method of applying evidence theory is through the utilization of the Dempster-Shafer rules. For a more comprehensive discussion, please refer to the following source: [12, 13].

Consider a finite set comprising elements denoted as:

$$\Omega = \{\omega_1, \omega_2, \omega_3, ..., \omega_n\} \tag{1}$$

called the *frame of discernment* (FoD). This collection represents mutually exclusive elements of a single cardinality, called *singletons.*

A *basic belief assignment* (BBA) or a piece of evidence is characterized by a function (mass functions) $m : 2^{\Omega} \rightarrow [0, 1]$ such that:

$$m(\emptyset) = 0 \tag{2}$$

$$\sum_{A \subseteq \Omega} m(A) = 1 \tag{3}$$

The quantity $m(A)$, known as the mass function, measures the credibility that one adheres exactly to hypothesis A; and (2) serving as an indicator of the closed world assumption [14]. If $m(A) > 0$, A represents a *focal element* of m.

Considering a basic belief assignment m, two notions can be disclosed, *a credibility or a belief function* (Bel) and a *plausibility function* (Pl) using the following expressions:

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{4}$$

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - Bel(\bar{A}) \tag{5}$$

$Bel(A)$ can be understood as the degree of total support to A, whereas $Pl(A)$ represents the extent to which doubt is absent. A.

If the frame of discernment is represented by a single focal element that is Ω, then the BBA m is considered *vacuous* and characterizes total *ignorance.*

Two mass functions m_1 and m_2 representing independent pieces of evidence can be merged by Dempster's rule which is defined as follows:

$$(m_1 \oplus m_2)(A) = \frac{1}{1-k} \sum_{B \cap C = A} m_1(B) m_2(C) \tag{6}$$

For all $A \subseteq \Omega$, $A \neq \emptyset$, and $(m_1 \oplus m_2)(\emptyset) = 0$. The constant k represents the degree of conflict of the two BBAs and is expressed as:

$$k = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{7}$$

In this work, the previously established concepts from evidence theory are integrated into an evidential framework that is plugged into a deep learning lidar-camera cross-fusion architecture. The primary objective is to leverage the respective advantages of these two frameworks to realize scene segmentation. The evidence theory is thus combined with the streamlined neural network referred to as Lite-CF, described in the next section. Originally designed as a probabilistic model for road detection, Lite-CF generates probability distributions from logits through a softmax layer [15]. To improve the handling of prediction uncertainty, an evidential approach recreates the last part of the architecture by replacing the softmax layer and the associated logits with a belief theory-based approach.

3 Proposed method

To handle data from various sensors effectively, architectures have been devised, which integrate multiple fusion methods such as late fusion, deep fusion, early fusion, and others. In a prior study [15], a lidar-camera fusion approach named lite cross-fusion was introduced, employing a fully convolutional neural network for road detection [16]. This deep fusion network demonstrated superior performance compared to early or late fusion methods. Consequently, it was integrated into another work [15], resulting in a reduction of computation complexity by over 15%. The model is built upon an encoder-decoder architecture that employs dilated convolution to consolidate contextual details while preserving resolution. This road detection fusion network serves as the foundation for the research presented in this paper. On another hand, considering the combination between evidence theory and neural network, [17] proposes an evidential classifier with a distance to prototypes approach that replaces the softmax decision layer.

Considering the two previous methods, namely the aforementioned reduced cross-fusion road detection (Lite-CF), and the evidential classifiers, this paper introduces a fusion of these two approaches. The combined architecture, Lite CF-Evi is a combination designed for semantic segmentation tasks. The overall architecture of the evidential Lite-CF is given in Fig. 1: The system comprises an encoder-decoder network, an evidential formulation layer, and a decision-making unit. In the encoding segment, there exist two processing pipelines, each consisting of 13 layers: one for feeding the LiDAR input and the other for camera frames. At each layer, information from one modality is combined with the corresponding layer from the other modality through a trainable weighted sum operation ($*b_i$ and $*a_i$ respectively, where i is the layer number). These fusion weights are adaptable, allowing the fusion's position and its extent to be fixed by the data.

After the LiDAR and camera inputs are transformed into Basic Belief Assignments (BBAs) within the evidential formulation layer, decisions can be rendered concerning specific elements within the power set 2^{Ω}. In the context of semantic segmentation, this power set encompasses elements such as "road", "vehicle ", and "background" elements in the probabilistic version, and additionally "ignorance" in the evidential formulation. Consequently, the evidential approach enables having an imprecise class prediction.

3.1 Evidential Inference

The evidential formulation layer takes as its input, the feature maps generated by the decoding section. When the decoder reaches its maximum resolution, Basic Belief Assignments (BBAs) are generated by assessing the distances between the corresponding feature maps (i.e., L18 in Fig. 1) and prototypes. In this way, the prototypes are learnt naturally, in an automatic manner. The technique is called distance to prototypes and it can be described in three steps [18] as follows:

Step 1: Calculate the distance to prototype: Consider $\boldsymbol{x}$ to be a feature vector symbolizing features of a pixel to be labeled possibly as road ω_1, vehicle ω_2 or

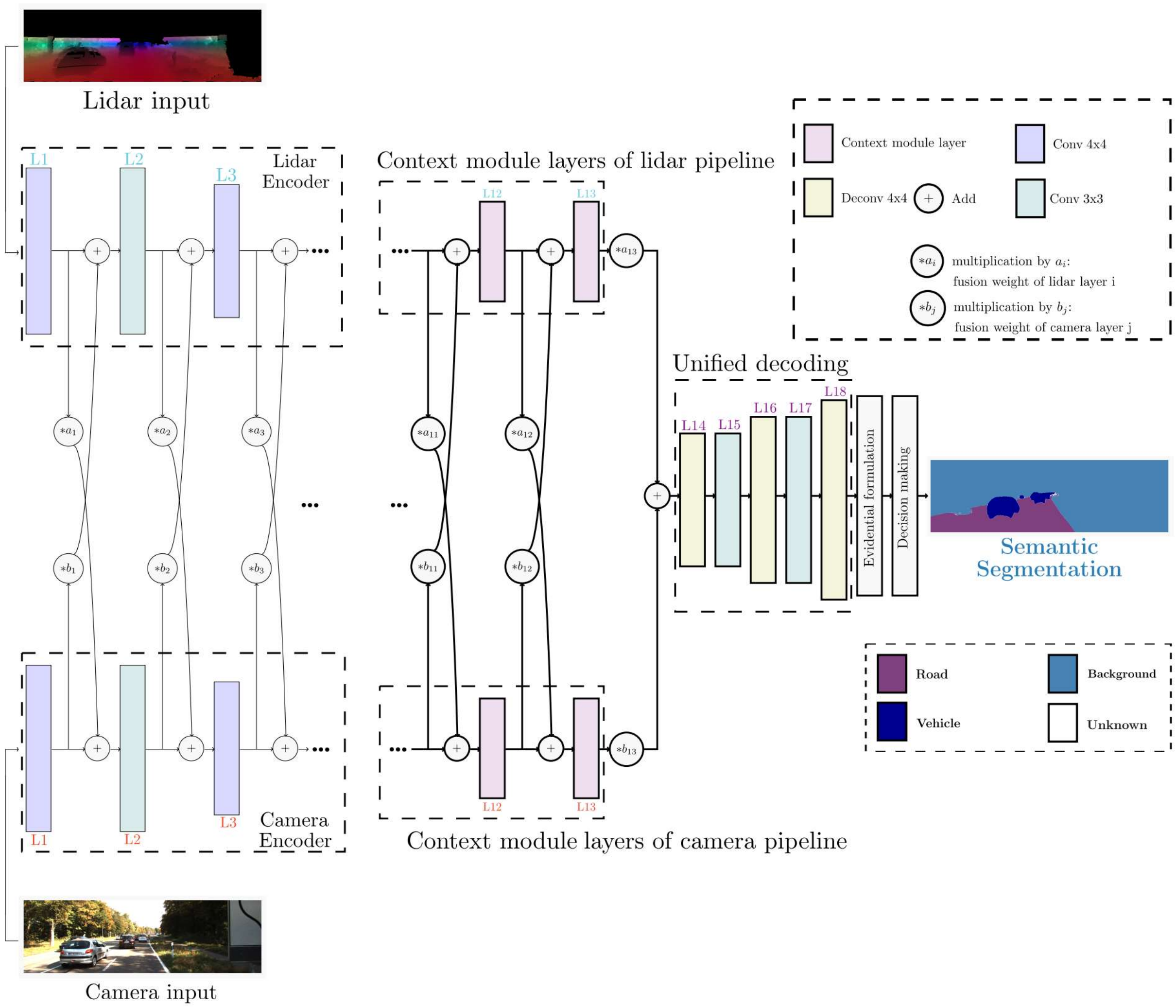


Fig. 1: Architecture Evidential Lite-CF (Lite CF-Evi)

background ω_3 (i.e., the FoD $\Omega = \{\omega_1, \omega_2, \omega_3\}$). The Euclidean distance d^i is realized between $\boldsymbol{x}$ and each prototype $\boldsymbol{p^i}$:

$$d^i = \|\boldsymbol{x} - \boldsymbol{p^i}\| \qquad i = 1, \cdots, n. \tag{8}$$

Step 2: Establish the correspondence of mass functions to prototypes and their interference: Each prototype $\boldsymbol{p^i}$ has a degree of membership u^i_j to each class ω_j, with a constraint $u^i_1 + u^i_2 + u^i_3 = 1$. Using the class membership u^i_j and the distance d^i, a BBA m^i is constructed as:

$$\begin{aligned} m^i(\{\omega_j\}) &= \alpha^i u^i_j \phi^i(d^i), \qquad j = 1, 2, 3 \\ m^i(\Omega) &= 1 - \alpha^i \phi^i(d^i), \end{aligned} \tag{9}$$

where $0 < \alpha^i < 1$ and the function ϕ^i is defined as:

$$\phi^i(d^i) = \exp\big(-\gamma^i (d^i)^2\big), \qquad \gamma^i > 0 \tag{10}$$

Step 3: DS combination rule: The mass functions from step 2 are merged using Dempster's rule (see (6)). The outcome combined with BBAs serves as the evidence for determining the pixel class.

The parameters linked to the prototype $\boldsymbol{p^i}$ (i.e., α^i, u^i_j, and γ^i), are intended to be incorporated into the evidential deep learning-based architectures as weighting factors. However, the learnable weights are not inherently restricted. Consequently, they are redefined and expressed in terms of certain real-numbered variables η^i, ξ^i, and β^i_j:

$$\gamma^i = (\eta^i)^2 \tag{11}$$

$$\alpha^i = \frac{1}{1 + \exp\{-\xi^i\}} \tag{12}$$

$$u^i_j = \frac{(\beta^i_j)^2 + \epsilon}{\sum\limits_{k=1}^{2} ((\beta^i_k)^2 + \epsilon)} \tag{13}$$

Equation (13) is slightly adjusted from the expression given in [18]. To prevent the membership values u^i_j from becoming zero, a small positive value denoted as ϵ is introduced. This precautionary measure is taken to limit the conflicts that could lead to Dempster's *total conflict*.

3.2 Decision making

After assessing the BBAs representing the evidence for each pixel, the ultimate objective is to decide the pixel classes. Initially, a decision based on interval distance is considered, as outlined in [19]. Here's an example with only 3 classes (in a semantic segmentation task: road, vehicle, and background):

Case i) The decision is constrained to singletons: The possible judgment elements are ω_1 (road), ω_2 (vehicle) and ω_3 (background). In this situation, the expression from [19] becomes:

$$\hat{X} = \arg \min_{X \in \{\omega_1, \omega_2, \omega_3\}} d_{BI}(m, m_X) \tag{14}$$

Case ii) The decision is not restricted: It could be interesting to consider assigning ambiguous pixels to less precise classes within Ω. The method can minimize classification errors by circumventing decisions that have more of an arbitrary nature.

4 Semantic segmentation results

4.1 Dataset

The aforementioned evidential cross-fusion model Lite-CF-Evi architecture is evaluated in the context of segmentation tasks against the KITTI semantic segmentation dataset. This dataset provides only 200 camera images similar to Stereo and Flow 2012/2015 datasets. Unfortunately, the dataset has no LiDAR frames. Consequently, the corresponding 3D point-cloud data for the camera images needs to be extracted from the big original raw dataset, as outlined in [20] with all the images from the KITTI dataset for all the tasks. Hence, 127 out of the 200 camera images have been successfully identified, along with their respective LiDAR frames. These LiDAR frames are subsequently projected and upsampled to create dense depth images. A 3D LiDAR point x is mapped into a point y in the camera plane according to the KITTI projection P, rectification R and translation T matrices:

$$y = P\,R\,T\,x \tag{15}$$

As the projected LiDAR scan is sparse, up-sampling is employed to generate a dense depth map, as depicted in Fig. 2. The up-sampling process is implemented following the method outlined in [15] and [21].

(a) Camera image

(b) Projected and up-sampled LiDAR points into the camera frame

Fig. 2: LiDAR pre-processing method

After the up-sampling process, the newly constructed dense depth images from LiDAR are integrated into the Lite-CF-Evi model in parallel with the camera images to feed the two pipeline inputs of the architecture.

Concerning the ground truth, the masks are simplified to 3 classes: road (magenta), vehicle (dark blue), and background (blue), according to the original annotation. The road class is preserved, however, the vehicle class incorporates car, truck, and bus annotations of the original ground truth. In turn, the background class encapsulates all the other classes, except for the above-mentioned ones. Figure 3a shows an illustration with an example of the original ground truth, while Figure 3b describes the simplified ground truth. The dataset consisted of 127 images: 114 for training and 13 for validation. This method has been exclusively assessed using the specially reconstructed KITTI semantic dataset, which includes the added LiDAR frames for the evidential cross-fusion architecture. To the best of the author's knowledge, this dataset has not yet been examined by any other methods, since LiDAR point clouds were included. The ground-truth masks are one-hot encoded and class weight is applied to address the unbalanced data. Consecutively, the model is trained for 500 epochs using mean squared error loss and Adam optimizer.

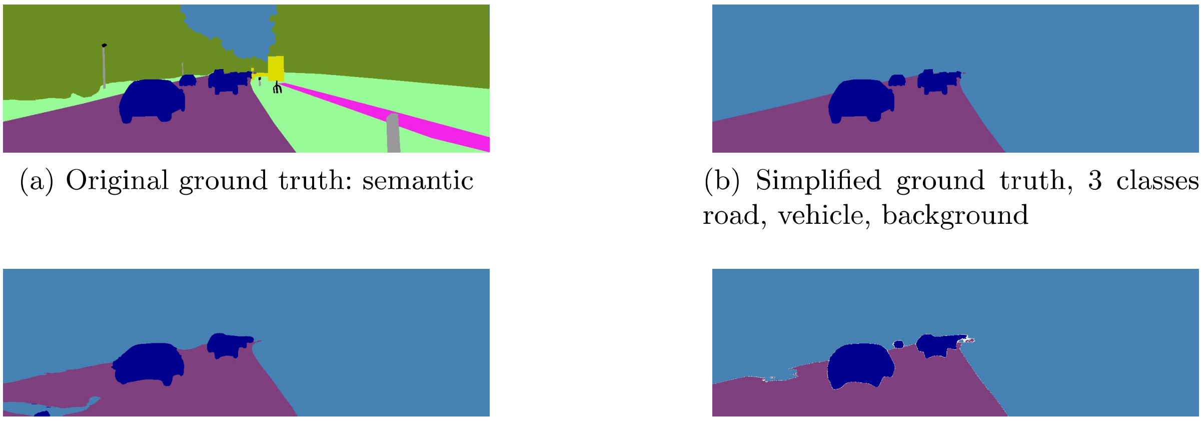
To measure the performances, the model is evaluated using the intersection-over-union metric, denoted as IoU, in accordance with the PASCAL VOC benchmark [22]:

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

with TP, FP, and FN, respectively, true positive, false positive, and false negative. The Lite-CF-Evi is evaluated for 3 classes in a probabilistic manner. It can be observed that the global mean IoU, 0.92707, in the evidential architecture is higher than 0.92384 for the probabilistic model. Individually over each class, the evidential model outperforms the probabilistic one (Table 4.1), and visually the results are better for the Lite-Cf Evi.

Table 4.1: Model performance comparison

Model arch.	mean IoU	mean IoU_{road}	mean $\text{IoU}_{\text{vehicle}}$	mean $\text{IoU}_{\text{background}}$
Probab. Lite-CF	0.92384	0.92713	0.87118	0.97322
Lite-Cf Evi	**0.92707**	**0.93163**	**0.87446**	**0.97513**



(a) Original ground truth: semantic

(b) Simplified ground truth, 3 classes: road, vehicle, background

(c) Probabilistic Prediction Lite-CF: road, vehicle, background

(d) Evidential Prediction with Lite-CF-Evi: road, vehicle, background, ignorance (white)

Fig. 3: Semantic segmentation results

One interesting part of the evidential formulation is that the decision-making can be adapted to derive from a fixed number of classes (equal to the number of singletons) to the maximum number of acts, $|2^{\Omega}| - 1$. However, often the desired decision elements are considered just the singletons, which are the main classes and the uncertainties. The third image below (Fig. 3c), on the left side represents the predicted image with the probabilistic model.

Finally, the second image from the right bottom part, Fig. 3d represents the predicted image with the evidential model (Lite-CF-Evi). It can be observed that classes road, vehicle, and background exhibit slightly higher accuracy in their predictions, with road class being notably precise. Furthermore, an additional class, denoted as "ignorance" (depicted in white), effectively captures pixels associated with uncertain predictions. This approach prevents the mis-classification of uncertain pixels into incorrect categories, a scenario that may arise when utilizing a probabilistic approach.

"Ignorance" primarily manifests itself at the class boundaries, where the model frequently provides errors in its predictions. Likewise, pixels from distant objects often lack sufficient information, suggesting that the model encounters challenges in classifying them due to data uncertainty. Consequently, these pixels are classified as "ignorance," offering improved comprehension and demonstrating the effectiveness of evidential reasoning in managing uncertainties.

5 Conclusion

In this paper, a camera-lidar fusion is proposed by using a deep learning architecture combined with evidence theory for intelligent vehicles perception. The combination is realized at the very last level, replacing the softmax decision with a decision based on distance to prototypes. The introduction of ignorance as a decision element further improves efficiency. Hence, distant points and ambiguous features can be categorized as “ignorance” rather than being erroneously assigned to specific predictions. Future directions involve enhancing the Lite CF-Evi model for various class configurations and more intricate tasks while maintaining the computational efficiency needed for real-time applications. Additionally, a more in-depth examination of the distribution and impact of “ignorance” is intended to be explored.

Acknowledgments
The authors extend their gratitude to the French National Research Agency (ANR) ANR JCJC EviDeep project, the University of Haute-Alsace, and the Pierre-et-Jeanne Spiegel Foundation for their support in the execution of this project.

References

1. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3) (2015) 211–252
2. Giurgi, D.V., Josso-Laurain, T., Devanne, M., Lauffenburger, J.P.: Real-time road detection implementation of unet architecture for autonomous driving. In: IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), IEEE (2022) 1–5
3. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241
4. Dempster, A.P.: A generalization of bayesian inference. Journal of the Royal Statistical Society: Series B (Methodological) **30**(2) (1968) 205–232
5. Beynon, M., Curry, B., Morgan, P.: The dempster–shafer theory of evidence: an alternative approach to multicriteria decision modelling. Omega **28**(1) (2000) 37–50
6. Laghmara, H., Boudali, M.T., Laurain, T., Ledy, J., Orjuela, R., Lauffenburger, J.P., Basset, M.: Obstacle avoidance, path planning and control for autonomous vehicles. In: 2019 IEEE Intelligent Vehicles Symposium (IV), IEEE (2019) 529–534
7. Xu, P., Davoine, F., Denœux, T.: Evidential combination of pedestrian detectors. In: British Machine Vision Conference. (2014) 1–14
8. Xu, P., Davoine, F., Bordes, J.B., Zhao, H., Denœux, T.: Multimodal information fusion for urban scene understanding. Machine Vision and Applications **27**(3) (2016) 331–349
9. Capellier, E., Davoine, F., Cherfaoui, V., Li, Y.: Evidential deep learning for arbitrary lidar object classification in the context of autonomous driving. In: 2019 IEEE Intelligent Vehicles Symposium (IV), IEEE (2019) 1304–1311
10. Denœux, T.: Logistic regression, neural networks and dempster-shafer theory: A new perspective. Knowledge-Based Systems **176** (2019) 54–67
11. Alhaija, H., Mustikovela, S., Mescheder, L., Geiger, A., Rother, C.: Augmented reality meets computer vision: Efficient data generation for urban driving scenes. International Journal of Computer Vision (IJCV) (2018)
12. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
13. Denœux, T., Dubois, D., Prade, H.: Representations of uncertainty in ai: beyond probability and possibility. In: A guided tour of artificial intelligence research. Springer (2020) 119–150

14. Martin, A., Jousselme, A.L., Osswald, C.: Conflict measure for the discounting operation on belief functions. In: 2008 11th International Conference on Information Fusion. (2008) 1–8
15. Geletu, M.N., Josso-Laurain, T., Devanne, M., Wogari, M.M., Lauffenburger, J.P.: Deep learning based architecture reduction on camera-lidar fusion for autonomous vehicles. In: 2nd International Conference on Computers and Automation (CompAuto), IEEE (2022) 25–31
16. Caltagirone, L., Bellone, M., Svensson, L., Wahde, M.: Lidar–camera fusion for road detection using fully convolutional neural networks. Robotics and Autonomous Systems **111** (2019) 125–131
17. Tong, Z., Xu, P., Denœux, T.: An evidential classifier based on dempster-shafer theory and deep learning. Neurocomputing **450** (aug 2021) 275–293
18. Denœux, T.: A neural network classifier based on dempster-shafer theory. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans **30**(2) (2000) 131–150
19. Dezert, J., Han, D., Tacnet, J.M., Carladous, S., Yang, Y.: Decision-making with belief interval distance. In: International conference on belief functions, Springer (2016) 66–74
20. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) (2013)
21. Premebida, C., Carreira, J., Batista, J., Nunes, U.: Pedestrian detection combining rgb and dense lidar data. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE (2014) 4112–4117
22. Everingham, M., Eslami, S., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1) (2015) 98–136