Evidential Deep Learning For Sensor Fusion

Mihreteab Negash Geletu[†‡], Jean-Philippe Lauffenburger[†], Thomas Josso-Laurain[†], Maxime Devanne[†]
and Mengesha Mamo Wogari[‡]
[†]*IRIMAS EA7499,* [‡]*AAiT-SECE*
[†]*Université de Haute-Alsace,* [‡]*Addis Ababa University*
[†]Mulhouse, France; [‡]Addis Ababa, Ethiopia
Email: mihreteab.negash@aau.edu.et, mengesha.mamo@aau.edu.et, and [†]firstname.name@uha.fr

***Abstract*—Perception in autonomous vehicles (AVs) is a challenging task. On the one hand, the driving environment is cluttered and the weather conditions vary. On the other hand, perception sensors have their own inherent shortcomings. To overcome these problems, deep learning-based methods often relying on probabilities are used. In this paper, deep learning-based multi-modal fusion architectures are implemented with the evidence theory. The evidential implementation defines the pieces of evidence using distances to prototypes of feature vectors obtained thanks to the neural networks. These belief functions are combined by Dempster's rule. Experimental analysis is done on the KITTI dataset and the analysis shows that the evidential models have better performance than the probabilistic baseline.**

***Index Terms*—Belief Functions, Deep Learning, Autonomous Vehicles, Environment Perception.**

I. Introduction

This paper has the objective of combining evidence theory with deep learning networks for environment perception in the application scope of Autonomous Vehicles (AVs). Environment perception in AVs is a challenging task. There are occlusion and truncation of objects. The day-night time change, illumination change, or simply driving through a tunnel creates difficult scenario for perception modules.

AVs are equipped with different perception sensors. Camera, LiDAR, radar and ultrasonic being the common types. However, these sensors have their own inherent shortcomings and a uni-modal approach may not be sufficient. Therefore, multi-modal sensor fusion is common in AVs [1]. For example, cameras are rich in resolution, and provide fine details and color information. Because of this, they are commonly used in perception tasks like classification, object detection and so on. However, they do not give depth information directly. Besides, their sensitivity to illumination variation is also problematic (e.g. entering or leaving a tunnel, and foggy or sunny day), and do not provide night-time vision. This limitation can be complemented by LiDAR sensors. Hence, multi-modal sensor fusion (e.g. fusing camera with LiDAR) is an important aspect in AV perception.

There are different multi-modal fusion architectures in perception tasks. Early fusion, late fusion and middle fusion are some to mention [1]. They are based on when the fusion happens in the processing pipeline of the perception algorithm. Accordingly, they have their own pros and cons on memory and computational power demand, complexity and flexibility. One example of a fusion architecture with extensive intermediate features combination is done in [2]. The fusion architecture is specifically called cross-fusion, which represents the fusion scheme and it fuses camera image and LiDAR scan.

Multi-modal fusion for environment perception tasks is currently implemented using Deep Learning (DL) due to its outstanding performance. DL has brought about a revolution in computer vision [3]. In DL-based perception networks, prediction scores are often represented using a softmax layer output as a probability distribution over a discrete set of elements, such as object classes in a classification problem. Even if the softmax layer is widely used in giving prediction outputs, it has a problem of inflating the probability of the predicted class because of the exponent employed. Moreover, probabilities cannot distinguish between the notion of conflict and ignorance in the uncertainty representation. These problems can be curbed by formulating deep learning-based models with an extended formalism for uncertainty handling like the evidence theory. Recently, works have been done to develop Evidential Deep Learning (EDL) networks mainly for image processing oriented applications [4, 5].

In this paper, starting from a state-of-art DL multi-modal framework, 3 optimized EDL fusion architectures are derived. Their aim is to fuse LiDAR point clouds with 2D camera images for road detection. It will be shown that these models provide better results than the original probability-based neural network with an improved computation cost. The rest of the paper is structured as follows: Section II covers a basic background description on evidence theory. A literature survey on EDL is presented in Section III. Section IV describes the evidential model development and experimental results are analyzed in Section V. Finally, Section VI concludes the paper.

II. Belief functions background

Evidence theory is a formalism for representing, reasoning and making decision with uncertainty [6]. It is also called Dempster-Shafer theory. The source of uncertainty could be randomness of some process. Such kind of uncertainty is called aleatory and it can not be reduced. Besides, uncertainty can rise from a lack of sufficient knowledge and it is then epistemic. This form of uncertainty can be reduced by acquiring additional information.

A. Basics in Evidence Theory

Let $\Omega = \{\omega_1, \omega_2, \omega_3, ..., \omega_n\}$ be the finite set of mutually exclusive elements, called the *Frame of Discernment* (FoD), and the mutually exclusive elements of single cardinality are called *singletons*. A *Basic Belief Assignment* (BBA), also called mass function, is a function $m : 2^\Omega \rightarrow [0, 1]$ such that

$$m(\emptyset) = 0 \tag{1}$$

$$\sum_{A \subseteq \Omega} m(A) = 1 \tag{2}$$

The BBA is called normalized (or, proper) when $m(\emptyset) = 0$ (i.e., a closed-world assumption). The quantity $m(A)$ measures the belief that one commits exactly to hypothesis A (i.e., the true answer to a certain question is in A), and it can not be assigned to any proper subset of A. If $m(A) > 0$, A is called a *focal set (or element)* of m. Given a BBA m, *belief function* Bel and *plausibility function* Pl are defined as

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{3}$$

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - Bel(\bar{A}) \tag{4}$$

$Bel(A)$ can be interpreted as the degree of total support to A, whereas $Pl(A)$ is the degree one fails to doubt A. A BBA with a single focal element is called *categorical* or *logical*, and if all the focal elements are singletons, it is called *Bayesian*. If the only focal element is Ω, the BBA is called *Vacuous* and represents total *ignorance* (i.e., completely non-informative).

Sources of evidence may not be reliable or their associated support can be inaccurate because they fail to take into account of some particular uncertainty affecting the evidence. In this situation discounting the support given by the mass function is relevant [6]. If $1 - \alpha$ is the degree of trust in the evidence as a whole, where $0 < \alpha < 1$, then α is the discount rate and the discounted mass function is given as

$$\begin{aligned} {}^{\alpha}m(A) &= (1 - \alpha)m(A) \quad \forall A \subset \Omega, \\ {}^{\alpha}m(\Omega) &= (1 - \alpha)m(\Omega) + \alpha. \end{aligned} \tag{5}$$

Discounting can be used to reduce the effect of sources which are not trusted in combining multiple sources of evidences. Otherwise, the strongest sources among conflicting ones can dominate the combination.

B. Distance to prototype BBAs

This section recalls how BBAs are defined and finally combined thanks to the distance to prototype method introduced by Denœux [4]. It will be further employed in Section IV to define the BBAs of the EDL neural networks developed. It follows a 3 steps approach:

1) Distance to prototype calculation: Let $\boldsymbol{x}$ be a feature vector representing features of an object to be classified possibly as ω_1 or ω_2 (i.e., the FoD $\Omega = \{\omega_1, \omega_2\}$). The Euclidean distance d^i is determined between $\boldsymbol{x}$ and each prototype $\boldsymbol{p^i}$:

$$d^i = \|\boldsymbol{x} - \boldsymbol{p^i}\| \qquad i = 1, \cdots, n, \tag{6}$$

where, n is the number of prototypes.

2) Basic belief assignment: Each prototype $\boldsymbol{p^i}$ has a degree of membership u^i_j to each class ω_j, with a constraint $u^i_1 + u^i_2 = 1$. Using the class membership u^i_j and the distance d^i, a BBA m^i is constructed as

$$\begin{aligned} m^i(\{\omega_j\}) &= \alpha^i u^i_j \phi^i(d^i)), \qquad j = 1, 2 \\ m^i(\Omega) &= 1 - \alpha^i \phi^i(d^i), \end{aligned} \tag{7}$$

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

$$\phi^i(d^i) = \exp(-\gamma^i (d^i)^2), \qquad \gamma^i > 0 \tag{8}$$

3) BBA combination: The BBAs constructed in step 2 are combined using Dempster's rule (see Section II-C).

The parameters associated with the prototype $\boldsymbol{p^i}$ (i.e., α^i, u^i_j and γ^i) are implemented in the evidentially formulated deep learning-based architectures as weights. Because of the constraints on these parameters, they are redefined in terms of some other variables η^i, ξ^i and β^i_j:

$$\gamma^i = (\eta^i)^2 \tag{9}$$

$$\alpha^i = \frac{1}{1 + \exp(-\xi^i)} \tag{10}$$

$$u^i_j = \frac{(\beta^i_j)^2 + \epsilon}{\sum_{k=1}^{2} ((\beta^i_k)^2 + \epsilon)} \tag{11}$$

Equation (11) is slightly modified from the expression given in [4]. A small positive number ϵ is introduced to avoid the membership values u^i_j from becoming zero. Otherwise, approximations in digital representation of numbers may cause the Dempster's rule to fail because of *total conflict*.

C. Evidence Combination

Two BBAs m_1 and m_2 representing independent pieces of evidence on a common frame Ω can be combined by Dempster's rule defined as

$$(m_1 \oplus m_2)(A) = \frac{1}{1 - k} \sum_{B \cap C = A} m_1(B) m_2(C) \tag{12}$$

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

$$k = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{13}$$

If $k = 0$, the two pieces of evidence m_1 and m_2 are said to be non-conflicting (i.e., each focal set of m_1 intersects all focal sets of m_2). If $k = 1$, the pieces of evidence are logically contradictory (i.e., total conflict). Thus, they cannot be combined by Dempster's rule. This rule is also called Dempster-Shafer (DS) rule of combination. The DS rule is commutative and associative. Therefore, in combining multiple sources of evidence, the combination order does not matter. There are also alternative combination rules like the *unnormalized* or *conjunctive* rule [7] which can identify partial conflicts.

D. Belief Interval Distance-based Decision

Given a BBA that represents a piece of evidence, the decision maker has the problem of choosing a particular element of the FoD that solves the problem under consideration. There are multiple ways to make decision. Classically, a decision can be made based on the value of belief Bel, plausibility Pl and probabilistic transformation [8]. However, they do not use the complete information represented by the BBAs. Besides, the probability transformations have their own drawbacks. These transformations lose all the exhaustive description of a situation provided by the evidential theory. A general decision rule that uses the whole information contained in the BBAs is proposed in [9]. It is based on belief interval BI which is defined as $BI = [Bl(X), Pl(X)]$ for $X \in 2^{\Omega}$. The rule uses a distance called belief interval distance d_{BI} between two BBAs m_1 and m_2 [10].

$$d_{BI}(m_1, m_2) = \sqrt{N_c \cdot \sum_{X \in 2^{\Omega}} d_W^2(BI_1(X), BI_2(X))}, \quad (14)$$

Where $N_c = 1/2^{n-1}$ and $d_W(BI_1(X), BI_2(X))$ is called the Wassertein's distance which is given as

$$d_W([a_1, b_1], [a_2, b_2]) = \sqrt{\left[\frac{a_1 + b_1}{2} - \frac{a_2 + b_2}{2}\right]^2 + \frac{1}{3}\left[\frac{b_1 - a_1}{2} - \frac{b_2 - a_2}{2}\right]^2}, \quad (15)$$

Where $[a_1, b_1]$ and $[a_2, b_2]$ are $BI_1(X)$ and $BI_2(X)$, respectively, associated with the BBAs m_1 and m_2. Given a BBA m, the decision $\hat{X}$ will be made by

$$\hat{X} = arg \min_{X \in 2^{\Omega} \setminus \emptyset} d_{BI}(m, m_X), \quad (16)$$

Where m_X is a *categorical* BBA focused on X. The argmin can be constrained only to some desired elements of the powerset 2^{Ω}. In this case, the expression in (16) becomes $\hat{X} = arg \min_{X \in 2^{\Omega} \setminus \emptyset \ s.t. \ c(X)} d_{BI}(m, m_X)$, where $c(X)$ is the constraint.

III. Related Work

The line of research in combining evidence theory with deep learning is based on a work on evidential k-nearest neighbor classification [11]. On top of that, in [4], evidence theory and neural networks are linked to form a prototype-oriented unified architecture where the prototypes are used to formulate basic belief functions.

Recently, this topic has gained some interest and gives rise to evidential deep learning networks in classification and semantic segmentation tasks.

A. Evidential Classification

In [12], an evidential neural network classifier is done on multiple types of classification tasks. The network architecture has CNN-based feature extractor layers, evidence theory-based belief function construction layer, and a decision making layer that uses the expected utility of assigning instances to possible classes in the frame of discernment (FoD) or in subsets. The work has investigated precise classification (i.e., assigning an instance to a single element of the FoD), set-valued classification (i.e., assigning an instance to a set of multiple elements of the FoD), and novelty (or outlier) detection on different tasks: image classification, signal classification, and semantic-relationship classification. The results show that evidential models have improved performance compared to the probabilistic models. A similar development is also used in [13] to build an evidential model providing a reduced error rate by rejecting ambiguous patterns.

For fast and automatic Covid-19 detection from computed tomography (CT) images, the evidential approach is used in [14]. It is composed of a CNN feature extractor and an evidential formulation and classification output parts. Its detection output is a binary classification as Covid-19 case or not. Since the annotated dataset is limited, a semi-supervised training mechanism is proposed. The model is reported to have higher accuracy than its probabilistic counterpart.

Besides the individual evidential classifiers, deep learning-based evidential classifier fusion has also shown an interesting result in [15]. It uses multiple evidential classifiers which are pre-trained on heterogeneous datasets of different class granularity. Experiments are conducted on fusing three evidential classifiers which are pre-trained on three datasets of objects and animals (birds, cats and dogs). The proposed evidential fusion architecture performs at least as good as the individual classifiers on their respective dataset and shows improvements on some classes, which have a joint support from the heterogeneous datasets.

B. Evidential Semantic Segmentation

Similarly to [12], Tong et al. proposed prototype-based evidential neural networks for semantic segmentation in [16]. A Fully Convolutional Network (FCN) is used for the image feature extraction. The evidential layer uses the same principles as the works above (distance to prototype BBAs and Dempster fusion). The proposed approach slightly improves precise segmentation accuracy. Besides, evidential imprecise segmentation gives high performance as it has capacity to assign ambiguous or less informative features into multi-class assignment rather than miss-classifying them. It also takes the advantage of being trained with soft labels.

In [17], an evidential architecture is proposed to segment Positron Emission Tomography - Computed Tomography (PET/CET) images. While the evidential part remains unchanged, the architecture has a U-Net CNN encoder-decoder feature extraction module. A loss function with two parts is used to train the network. One part determines the segmentation accuracy and the other quantifies the segmentation uncertainty. The proposed model is evaluated using multiple metrics (dice score, sensitivity, specificity, precision and F1 score), and it outperforms the softmax-based U-Net and other related probabilistic models.

Huang et al. developed a close approach in [18] using two models. The first, called ENN-UNet, uses the prototype-based evidential neural network in [4], and the second named RBF-UNet is based on the re-interpretation of generalized logistic regressions in [5]. The latter is implemented with an heuristic approach employing a radial basis function network and the two approaches are reported to have similar performance.

In another work, an evidential PET and CT fusion is proposed for lymphoma segmentation [19]. It uses two uni-modal models trained separately to get segmentation maps, which are evidentially fused. The uni-modal models output probabilistic maps are considered as Bayesian mass functions to be combined by DS rule. The fusion model is trained end-to-end with a loss function that has three parts to reduce the loss in the two uni-modal segmentation maps and the final fused map. The proposed fusion model is reported to have better performance than uni-modal approaches and other state-of-the-art methods.

C. Discussion

The above literature review shows that the evidential classifiers and semantic segmentation deep learning models primarily rely on prototypes to generate BBAs, or consider prediction probabilities as Bayesian mass functions. The pieces of evidence represented by BBAs are finally combined using Dempster's rule. Generally, the evidential models have shown superior performance than the corresponding probabilistic models in precise prediction, imprecise prediction, confidence calibration, novelty (or outlier) rejection and sensor fusion. Their decision making rule is commonly based on belief, plausibility or probabilistic transformation. Besides, some decision rules that require optimization for parameter estimation are also used, and these have some development complexity. The evaluation is based on conventional metrics and self defined metrics. These works do not consider the fusion of heterogeneous data types issued from real-field sensors.

In this work, the objective is to extend the prototype-based BBAs generation to the case of multimodal fusion for environment perception. It takes into account the computational power constraint of the application domain, AVs, by considering EDL complexity reduction.

IV. Evidential Deep Learning Multimodal Fusion

The state-of-the-art above shows recent works primarily on prototype-based EDL, mainly devoted to medical as well as animal and object classification applications. However, such approaches mixing model-driven and data-driven methods are of great interest and some works have already be done in robotics and intelligent vehicles [20, 21]. This section describes prototype-based evidential deep-learning architectures for robot environment perception derived from the literature described in Section III. These architectures are based on a multimodal (multisensor) deep learning cross-fusion (CF) scheme proposed in [2].

A. Multimodal Cross Fusion

In [2], Caltagirone et al. proposed an encoder-decoder architecture considering two complementary input sensors (a 2D-camera and a 3D-LiDAR) for road detection. This architecture has two processing pipelines, one per modality, composed of an encoder (5 layers), a context module (9 layers) and a decoder (6 layers). A classical softmax layer gives the probabilistic output of the road detection. A noticeable particularity of this network is the Cross Fusion (CF) between the pipelines: each layer of one modality is fused with the corresponding layer of the other modality by a weighted sum operation. In order to guarantee its real-time implementation, the CF has firstly been optimized by Geletu et al. [22] who proposed 2 models: a cross fusion architecture with a simplified decoder and a further optimized one with a reduced backbone. These architectures are reported to have close performance, smaller size and lower runtime than the original Cross Fusion. However, all the models are probabilistic. A preliminary work that leverages one of the optimized model is implemented evidentially with experiments emphasized on its applicability in AVs [23]. This paper shows how the 3 models (i.e., the optimized and original) have been adapted to the evidential framework with emphasis on the fusion step as described below. It also provides a quantitative performance evaluation of these models using a real sensor dataset.

B. Evidential Cross Fusion (ECF)

The evidential formulation of CF follows the work recalled in Section II-B. After the BBA combination, a decision is made using the belief interval distance d_{BI} (see II-D). Fig. 1 gives a generic architecture of the evidential formulation. As can be seen from the figure, features are extracted from the encoder-decoder-based architecture. The layer $L1$ in the evidential formulation computes the distance to prototypes of a feature vector, and activates it. BBAs associated with prototypes are then generated in layer $L2$. For its evidential implementation (ECF), the probabilistic logits in the CF are removed and the layers before logits are weighted sum to form the feature maps, which are fed into the evidential formulation. Finally d_{BI} is employed to make decision based on the combined BBA.

C. Evidential Cross Fusion with a Unique Decoder (ECFU)

The Cross Fusion network with a reduced decoder is a reduction of the CF network with comparable performance [22]. The reduction follows an analysis on the CF fusion weights, which dictates that the fusion between the two processing pipelines is not favoured at the decoder section. As a result, a single unified decoder which reconstructs from a fused latent space representation is employed. Though there is no fusion in the decoder section, it has the same number of layers as that of CF. The computational space gained by the reduction is leveraged to augment evidence theory without incurring computational burden. In the evidential implementation ECFU, the layers before logit of the corresponding probabilistic

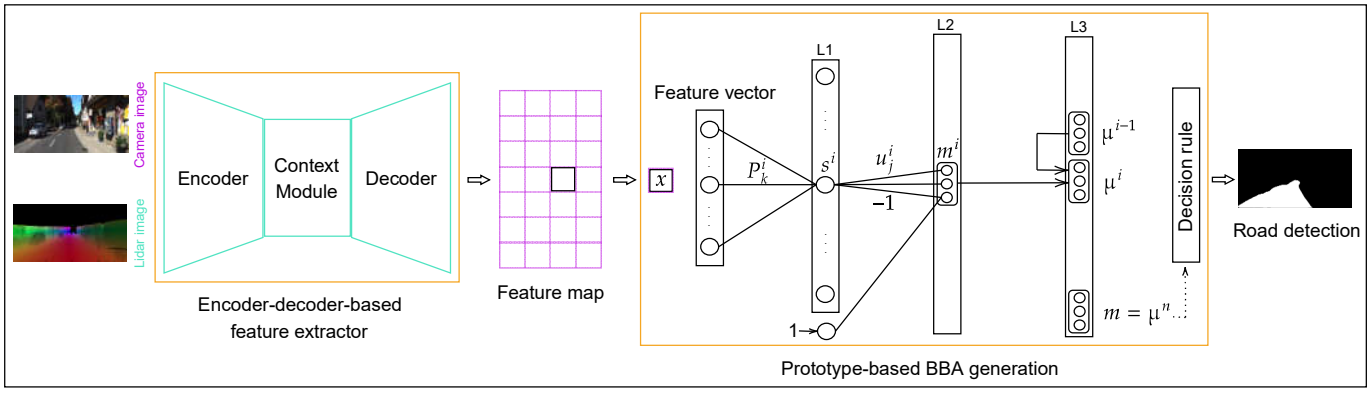

Fig. 1. Generic architecture of evidential formulation, $s^i = \alpha\phi^i(d^i)$ is the activation and $m = \mu^n$ is the combined BBA.

model is used as feature maps. Based on this features, BBAs generation and decision making follows as shown in Fig. 1.

D. Evidential Light Cross Fusion (ELCFU)

The Evidential Light Cross Fusion architecture is an optimization of the ECFU. It reduces the network context module by removing a layer [22]. This yields a further gain in the execution speed and model size. The ELCFU has a total of 19 layers. Its evidential implementation also removes the logits and uses the layer before for the three steps evidential formulation procedure depicted in Fig. 1.

V. Experimental Results

The prediction outputs for the three evidential models ECF, ECFU and ELCFU are evaluated using a set of evaluation metrics. They are also compared with the original probabilistic model (CF). To make fair comparison and avoid biases, specific details in dataset pre-processing, model development, training and evaluation are detailed.

A. Dataset and Pre-processing

The KITTI dataset on the task of road detection [24] is used in this work. The dataset gives camera images and LiDAR scans among other sensors measurements. The 3D point cloud from the LiDAR is projected and up-sampled to get camera like dense depth image [22]. Fig. 2 shows a camera image and its corresponding projected and up-sample LiDAR depth map. The dataset precisely labels pixels in the ground truth as not-road and road. This information is encoded into categorical BBAs. Of course, obviously, there is an avoided imprecision and uncertainty in the KITTI ground truth labeling. For example, ambiguities in road area which is underneath a car or leaves covering the road [24]. However, in this work the ground truth labeling is considered to be precise and certain.

B. Training Details and Evaluation Scheme

The number of prototypes is set to 6 from preliminary experimental observations. The high level per pixel feature vector has a dimension of $1 \times 1 \times 8$, which is the same as in the CF architecture [2]. The loss regularization factor of the

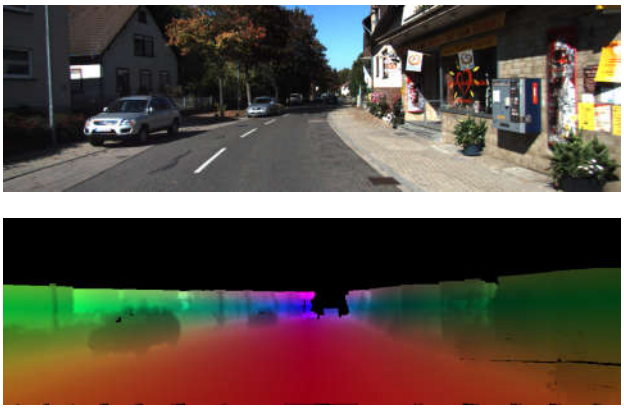

Fig. 2. Dataset pre-processing: projected and up-sampled LiDAR depth map and corresponding camera image.

parameter α^i associated with the prototypes $\boldsymbol{p^i}$ is set to 0.01 similarly to [4, 14]. The training dataset has 289 frames. These frames are split according to stratified 10-fold cross validation scheme into training and validation splits. The validation split is used to choose the best model in the training. Since the dataset size is small, an augmentation by a random rotation of the images in range of $[-20^o, 20^o]$ about the center is used. A summary of the model hyper-parameters and the training details are given in Table I.

TABLE I
Model hyper-parameters and training details

Description	Setting
# Prototypes	6
High level feature vector size	$1 \times 1 \times 8$
Regularization factor (loss function)	0.01
Input size (HxW)	384x1248
Augmentation	Random rotation
Optimization	Adam
Loss	Mean squared error
Learning rate	Polynomial decay
Epoch	200
Batch size	1

The KITTI evaluation metrics maximum F1 (MaxF), pre-

cision (PRE) and recall (REC) are used to evaluate the models [24]. In addition, the error rate (ER) is employed, which is also utilized in similar works [4, 5]. Furthermore, the frame per second (FPS) is calculated by averaging the runtime over the 289 training frames on an Nvidia RTX 2070 GPU board. The evaluation is made according to a stratified 10-fold cross-validation, 10 separate trainings and evaluations are made for each model. Therefore, results are expressed as mean $\mu \pm$ and standard deviation σ. The d_{BI} decision rule in equation (16) can be constrained to singletons. Therefore, the decision $\hat{X}$ is made among the possible decision elements ω_1 and ω_2:

$$\hat{X} = arg \min_{X \in \{\omega_1, \omega_2\}} d_{BI}(m, m_X), \tag{17}$$

Since preliminary experimental results showed that the direct implementation of the d_{BI} rule (i.e., (17)) is computationally expensive, it is algebraically simplified to a mathematically equivalent expression for the evaluation:

$$\hat{X} = arg \max_{X \in \{\omega_1, \omega_2\}} m(X) \tag{18}$$

C. Performance Evaluation

Table II shows the performance comparison among the probabilistic model (CF) and the evidential formulations (ECF, ECFU and ELCFU). As can be seen, the evidential models have better performance than the probabilistic CF. ECF, the evidential evolution of the original cross-fusion has the highest value in MaxF, PRE, REC and the lowest in ER. The evidential formulation also results in a decrease in the number of model parameters. Because, reduced deep learning architectures are employed and further, the logits layer has been removed (see Section IV). Considering the model optimization, ELCFU has around 16% parameter reduction but provides better performance than the CF. These reductions and the simplification in the decision rule also make the evidential models fast. In terms of real-time execution, ELCFU has again superior performance compared to the CF. Its FPS is increased by about 22%. This is an interesting result that an evidential model, which leverages architectural reduction, has gained performance while being faster and having fewer number of parameters than the probabilistic baseline. Furthermore, in the performance evaluation metrics MaxF, PRE, REC and ER, the statistical variation of the evidential results is small compared to the CF.

A sample heat map from the ELCFU model is given in Fig. 3 to visualize the underlying BBAs in the prediction. Basically these BBAs are used in the decision making. Therefore, a collective analysis on them is relevant. As can be seen from the figure, BBAs have three focal elements (i.e., not-road, road and ignorance) though the training used precise labels (i.e. support to either not-road or road, not both). It can also be observed that the not-road and road areas generally have strong support (i.e., high mass values). The mass of ignorance, on the other hand, is small (≤ 0.08) in the entire area. This shows that the evidential model has learned well

TABLE II
MODEL PERFORMANCE COMPARISON [MAXF, PRE, REC AND ER ARE IN PERCENTAGE]

Model arch.	# model param.	MaxF	PRE	REC	ER	FPS
CF	3,246,830	96.25 ± 0.71	96.46 ± 0.66	96.05 ± 1.06	1.34 ± 0.26	27
ECF	3,246,462	**97.08 ± 0.43**	**96.84 ± 0.62**	**97.33 ± 0.66**	**1.05 ± 0.16**	26
ECFU	3,032,236	96.97 ± 0.48	96.71 ± 0.55	97.23 ± 0.74	1.09 ± 0.18	31
ELCFU	**2,737,018**	96.91 ± 0.36	96.74 ± 0.56	97.09 ± 0.71	1.11 ± 0.14	**33**

that it extracts relevant features to classify pixels into one of the two classes. Therefore, it avoids non-informativeness by maintaining small ignorance values. Even if the ignorance level is small, it is discriminated among three regions: road, road edge and not-road with decreasing in values. The road edge has also comparatively lower support to either one of the two classes. This is as expected, because, the area is ambiguous. Therefore, it can be seen that the evidential models have learned well with a highly reduced non-informativeness. Hence, given a frame from the dataset, the evidential models try to segment it precisely rather than giving a handful of ignorance prediction.

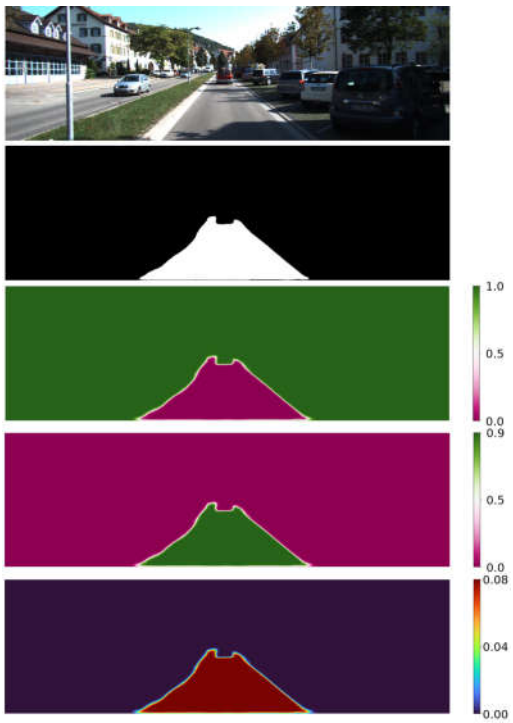

Fig. 3. Sample heat map of BBAs. From top to bottom: camera image, model evaluation result (black: not-road, white: road), heat map of $m(\omega_1)$ (not-road), heat map of $m(\omega_2)$ (road), and heat map of $m(\Omega)$ (ignorance).

VI. Conclusion

Evidential deep learning-based multi-modal perception networks are implemented in this paper. Feature vectors defined by the neural networks are converted into BBAs using a distance-to-prototype specific layer. The BBAs are fused by Dempster's rule and a belief interval distance-based decision is performed. A particular interest was paid on the computational constraints of the Autonomous Vehicles application domain by optimizing the multi-modal fusion architectures. Applied on a road detection application considering camera images and sparse LiDAR data, it is shown that the evidential models have better performance than the probabilistic baseline with standard metrics (precision, recall, error rate, etc.). Besides, their model complexity is lower, attaining a model parameter reduction of about 16%. Furthermore, they generally have a fast runtime. Particularly, an increase in the frame per second of more than 22% compared with the baseline is achieved with the best optimized structure. In addition, these models have small statistical dispersion. In the future, this result can be further extended with the analysis of imprecise prediction, outlier detection and confidence calibration. Besides, alternative decision rules can also be investigated.

Acknowledgment

The authors gratefully acknowledge the financial support from Ethiopian Ministry of Education and Embassy of France in Ethiopia and to the African Union under the Ethio-French PhD scholarships program in engineering.

References

[1] D. Feng et al., "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[2] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, "Lidar–camera fusion for road detection using fully convolutional neural networks," *Robotics and Autonomous Systems*, vol. 111, pp. 125–131, 2019.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] T. Denoeux, "A neural network classifier based on dempster-shafer theory," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 30, no. 2, pp. 131–150, 2000.

[5] T. Denœux, "Logistic regression, neural networks and dempster–shafer theory: A new perspective," *Knowledge-Based Systems*, vol. 176, pp. 54–67, 2019.

[6] G. Shafer, *A mathematical theory of evidence*. Princeton University Press, 1976.

[7] P. Smets, "The combination of evidence in the transferable belief model," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 12, no. 5, pp. 447–458, 1990.

[8] ——, "Decision making in a context where uncertainty is represented by belief functions," in *Belief functions in business decisions*. Springer, 2002, pp. 17–61.

[9] J. Dezert, D. Han, J.-M. Tacnet, S. Carladous, and Y. Yang, "Decision-making with belief interval distance," in *International conference on belief functions*. Springer, 2016, pp. 66–74.

[10] D. Han, J. Dezert, and Y. Yang, "Belief interval-based distance measures in the theory of belief functions," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 6, pp. 833–850, 2016.

[11] T. Denoeux, "A k-nearest neighbor classification rule based on dempster-shafer theory," *IEEE transactions on systems, man, and cybernetics*, vol. 25, no. 5, pp. 804–813, 1995.

[12] Z. Tong, P. Xu, and T. Denoeux, "An evidential classifier based on dempster-shafer theory and deep learning," *Neurocomputing*, vol. 450, pp. 275–293, 2021.

[13] ——, "Convnet and dempster-shafer theory for object recognition," in *International Conference on Scalable Uncertainty Management*. Springer, 2019, pp. 368–381.

[14] L. Huang, S. Ruan, and T. Denoeux, "Covid-19 classification with deep neural network and belief functions," in *The Fifth International Conference on Biological Information and Biomedical Engineering*, 2021, pp. 1–4.

[15] Z. Tong, P. Xu, and T. Denœux, "Fusion of evidential cnn classifiers for image classification," in *International Conference on Belief Functions*. Springer, 2021, pp. 168–176.

[16] Z. Tong, P. Xu, and T. Denoeux, "Evidential fully convolutional network for semantic segmentation," *Applied Intelligence*, vol. 51, no. 9, pp. 6376–6399, 2021.

[17] L. Huang, S. Ruan, P. Decazes, and T. Denoeux, "Evidential segmentation of 3d pet/ct images," in *International Conference on Belief Functions*. Springer, 2021, pp. 159–167.

[18] L. Huang, S. Ruan, P. Decazes, and T. Denœux, "Lymphoma segmentation from 3d pet-ct images using a deep evidential network," *International Journal of Approximate Reasoning*, vol. 149, pp. 39–60, 2022.

[19] L. Huang, T. Denœux, D. Tonnelet, P. Decazes, and S. Ruan, "Deep pet/ct fusion with dempster-shafer theory for lymphoma segmentation," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2021, pp. 30–39.

[20] E. Capellier, F. Davoine, V. Cherfaoui, and Y. Li, "Fusion of neural networks, for lidar-based evidential road mapping," *Journal of Field Robotics*, vol. 38, no. 5, pp. 727–758, 2021.

[21] X. Yu, G. Franchi, J. Gu, and E. Aldea, "Discretization-induced dirichlet posterior for robust uncertainty quantification on regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 6835–6843.

[22] M. N. Geletu, T. Josso-Laurain, M. Devanne, M. M. Wogari, and J.-P. Lauffenburger, "Deep learning based

architecture reduction on camera-lidar fusion for autonomous vehicles," in *2022 2nd International Conference on Computers and Automation (CompAuto)*. IEEE, 2022, pp. 25–31.

[23] M. N. Geletu et al., "Evidential deep learning-based multi-modal environment perception for intelligent vehicles," in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023, pp. 1–6.

[24] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*. IEEE, 2013, pp. 1693–1700.