# Perception multimodale de l'environnement basée sur l'apprentissage profond évidentiel pour les véhicules intelligents

Vasile Giurgi, Mihreteab Negash Geletu, Thomas Josso-Laurain, Maxime Devanne, Jean-Philippe Lauffenburger, Mengesha Wogari

# Perception multimodale de l'environnement basée sur l'apprentissage profond évidentiel pour les véhicules intelligents

D-V. Giurgi[*,1]
T. Josso-Laurain[1]
J-P. Lauffenburger[1]
M. N. Geletu[*,1,2]
M. Devanne[1]
M. M. Wogari[2]

[1] IRIMAS-UR7499, Université de Haute-Alsace, Mulhouse, France
[2] AAiT-SECE, Université d'Addis-Abeba, Addis-Abeba, Éthiopie

vasile.giurgi@uha.fr
[*]Contribution égale

## Résumé

*Les laboratoires de recherche et les entreprises s'intéressent aux véhicules intelligents (VI) afin de révolutionner les systèmes de transport. Étant donné que l'environnement de conduite peut être encombré et que les conditions météorologiques peuvent varier, la perception de l'environnement dans les VI représente une tâche difficile. C'est pourquoi des capteurs multimodaux sont utilisés. De plus, en matière de perception, des performances exceptionnelles sont obtenues en utilisant des algorithmes d'apprentissage profond qui reposent sur des probabilités. Ces dernieres ne permettent néanmoins pas une représentation de toutes les imperfections des données. Pour contourner ce problème, dans ce travail, la théorie de l'évidence est combinée avec une architecture d'apprentissage profond destinée à fusionner les données d'une caméra et d'un lidar 3D. Le couplage est basé sur la génération de fonctions de croyance en utilisant la distance aux prototypes. Il utilise également une règle de décision basée sur la distance euclidienne. Comme les VI ont une puissance de calcul limitée, une architecture d'apprentissage profond réduite est choisie dans cette formulation. Dans la tâche de détection de routes, l'approche évidentielle surpasse l'approche probabiliste. En outre, les caractéristiques ambigües, telles que les régions imprécises entre les limites des classes segmentées, peuvent être prudemment considérées comme de l'ignorance plutôt que de prendre une décision aléatoire. Le couplage est également étendu à la tâche de segmentation sémantique. Cette extension montre que la formulation évidentielle peut être facilement adaptée au cas multiclasses. Par conséquent, la formulation évidentielle est générique et produit une prédiction plus pertinente. Plus précisément, il prédit comme une certaine classe uniquement la région de pixels où le modèle est confiant à ce sujet, évitant d'attribuer les données d'imprécision (régions où le modèle n'est pas confiant) à la classe. La formulation évidentielle est aussi polyvalente tout en maintenant le compromis entre les performances et le coût de calcul dans les VI. Ce travail utilise le dataset KITTI.*

## Mots Clef

véhicules intelligents, perception de l'environnement, théorie de l'évidence, apprentissage profond.

## Abstract

*Intelligent vehicles (IVs) are pursued in both research laboratories and industries to revolutionize transportation systems. Since the driving surroundings can be cluttered and the weather conditions may vary, environment perception in IVs represents a challenging task. Therefore, multimodal sensors are engaged. Furthermore, in perception, outstanding performance is obtained by employing deep learning algorithms which rely on probabilities and therefore are not suitable to represent data uncertainties. To circumvent this, in this work, evidence theory is combined with a camera-lidar-based deep learning fusion architecture. The coupling is based on generating basic belief functions using distance to prototypes. It also uses a distance-based decision rule. Because IVs have constrained computational power, a reduced deep learning architecture is chosen in this formulation. In the task of road detection, the evidential approach outperforms the probabilistic one. Besides, ambiguous features can be prudently set as ignorance rather than making a random decision using probability. The coupling is also extended to the task of semantic segmentation. The extension shows that the evidential formulation can be easily adapted to the multi-class case. Therefore, the evidential formulation is generic and produces a more accurate and versatile prediction while maintaining the trade-off between performances and computational costs in IVs. This work uses the KITTI dataset.*

## Keywords

intelligent vehicles, environment perception, evidence theory, deep learning.

# 1 Introduction

La perception dans les véhicules intelligents (VI) est une tâche difficile. L'environnement de conduite peut être encombré et les conditions météorologiques peuvent varier. La perception est réalisée par un ensemble multimodal de capteurs tels que des caméras, des LiDAR, des radars, etc. Les entrées provenant de ces modalités sont traitées pour obtenir une compréhension de l'environnement de conduite, qui sera utilisée pour actionner le véhicule. Par conséquent, une défaillance de la perception se propagera pour déclencher une action de contrôle erronée. Des systèmes de perception performants sont donc nécessaires.

La perception s'appuie de plus en plus sur l'apprentissage profond qui a gagné en popularité depuis qu'une architecture de réseau neuronal convolutif profond appelée AlexNet a obtenu des performances nettement supérieures dans un défi de reconnaissance visuelle appelé ImageNet [1, 2]. Par conséquent, l'apprentissage profond a été utilisé pour différentes tâches de perception, telles que la classification, la détection d'objets, la segmentation sémantique, et d'autres encore. Dans cet algorithme, la sortie de prédiction repose souvent sur la fonction sigmoïde et la fonction softmax pour la prédiction binaire et multi-classes, respectivement. Ces sorties sont des probabilités définies sur un ensemble de classes de prédiction mutuellement exclusives. Elles ne peuvent pas distinguer l'absence d'informations des données contradictoires. Par exemple, dans le cas de la détection d'une route, la prédiction ne peut être attribuée qu'à la classe "route" ou "hors route" même s'il n'y a pas assez de preuves pour cette discrimination. L'approche est donc susceptible de prendre une mauvaise décision. Au contraire, il existe un meilleur formalisme bien établi de représentation enleverdes imperfections des données appelé théorie de l'évidence. Cette théorie, également connue sous le nom de fonctions de croyances (BF), est proposée pour la première fois par Dempster et Shafer [3] (théorie DS) pour représenter les éléments de preuve (croyances) pour les modèles incertains. Les principales caractéristiques de la théorie de l'évidence sont les suivantes : *généralité* (elle étend à la fois la logique propositionnelle et le raisonnement probabiliste), *opérationnalité* (fonctionne avec des éléments de preuve élémentaires couplés à la règle de Dempster-Shafer), et *scalabilité* (le raisonnement fondé sur des fonctions de croyance est utilisé pour résoudre des problèmes complexes), ce qui la rend plus complexe que la théorie des probabilités.

Dans les applications de navigation autonome, telles que l'évitement d'obstacles, les fonctions de croyance se sont avérées efficaces. Par exemple, les applications de l'environnement de perception, comme l'occupation de la grille d'un capteur LiDAR. Dans ce travail, le conflit est exprimé d'une manière plus représentative [4]. Dans le cadre des tâches de détection de piétons, les règles de combinaison évidentielles ont permis d'obtenir des performances considérables par rapport à l'approche bayésienne [5]. De plus, dans la perception multi-modale, la théorie évidentielle gère l'information manquante, l'imprécision et l'ignorance. Dans [6], les images de segmentation sémantique KITTI provenant de divers capteurs, caméras et différentes couches de LiDAR sont incorporées ensemble, ce qui permet d'élargir les classes d'objets ou le nombre de capteurs. L'approche permet d'améliorer les performances pour une meilleure compréhension de la zone carrossable. Dans le domaine de l'apprentissage profond, [7] propose une architecture de réseau neuronal basée sur le perceptron multicouche (MLP) pour classifier des objets LiDAR arbitraires pour la perception. Leur modèle remplace l'approche probabiliste par une méthode d'inférence évidentielle, inspirée du classifieur logistique généralisé de Denoeux [8]. Ainsi, les cadres basés sur les fonctions de croyance donnent des résultats prometteurs dans les systèmes de perception pour les tâches de segmentation de routes et de détection multi-objets, qui sont les principaux sujets abordés dans ce travail. Le but est de fournir un modèle d'apprentissage profond évidentiel qui fusionne les informations provenant de différents capteurs pour atteindre des capacités de conduite autonome.

La structure de l'article se poursuit avec les sections suivantes : État de l'art (méthodes de fusion de données et théorie de l'évidence), Méthode proposée (architecture de modèle d'apprentissage profond par fusion croisée couplée à l'inférence évidentielle), Résultats expérimentaux (traitement des données, décision, résultats), et Conclusion.

# 2 État de l'art

Un réseau de fusion multimodale basé sur l'apprentissage profond sera combiné à la théorie évidentielle pour mieux gérer l'incertitude de la prédiction dans la perception de l'environnement des VI. Cette section porte sur les aspects de l'apprentissage profond, de la fusion multimodale et de la théorie évidentielle.

## 2.1 Fusion lidar-caméra par apprentissage profond

Une fusion lidar-caméra appelée fusion croisée (CF) est proposée sur la base d'un réseau de neurones entièrement convolutif pour la détection des routes dans [9]. Cette fusion multimodale par réseaux de neurones a été appliquée à la détection de routes à partir du dataset KITTI et a obtenu de bons résultats [10, 11]. Elle est basée sur une architecture encodeur-décodeur, où la convolution dilatée est utilisée pour agréger les informations contextuelles sans perdre la résolution. Contrairement à la pré-fusion et à la post-fusion, où la position de la fusion est fixée respectivement au début et à la fin d'un pipeline de traitement, dans CF, la position et la zone de fusion sont apprises à partir des données de formation elles-mêmes. Il a été rapporté que la CF surpasse les pré et post fusions des entrées LiDAR et caméra [9]. Ce réseau de fusion constitue la base de notre travail. Du point de vue des performances en temps réel [12], la taille du réseau CF a été réduite dans [13], tout en maintenant des performances comparables. Le réseau ré-

duit est appelé Lite-CF et présente une réduction de plus de 15% des paramètres du modèle. La partie du Lite-CF utilisée dans ce travail est représentée sur la Fig. 1. L'illustration met en évidence les deux modules, respectivement caméra et LiDAR, et la fusion croisée entre eux. La plupart de l'architecture est la même que celle de la ligne de base mentionnée. Cependant, en ce qui concerne le module de contexte, certaines couches de la ligne de base originale sont remplacées par la formulation évidentielle. Les avantages de cette réduction sont pris en compte pour brancher la théorie évidentielle au lieu de la probabiliste, sans causer une perte significative de puissance de calcul.

### 2.2 Principes de base de la théorie évidentielle

La théorie de l'évidence est un formalisme permettant de raisonner et de prendre une décision dans l'incertitude.Une façon d'aborder la théorie de l'évidence est d'utiliser les règles de Dempster-Shafer basées sur les fonctions de croyance. Dans cette section, une brève introduction à la théorie est donnée. Une discussion détaillée peut être trouvée dans [14, 15].

Soit $\Omega = \{\omega_1, \omega_2, \omega_3, ..., \omega_n\}$ un ensemble fini d'éléments mutuellement exclusifs et exhaustsifs, appelé *cadre de discernement* (FoD), et les éléments mutuellement exclusifs de cardinalité unique sont appelés *singletons*. Une *assignation de croyance basique* (BBA) représente une fonction $m : 2^\Omega \rightarrow [0, 1]$ telle que :

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

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

La quantité $m(A)$, connue sous le nom de fonction de masse, mesure la croyance selon laquelle on s'engage exactement dans l'hypothèse A. Si $m(A) > 0$, $A$ est appelé un *élément focal* de $m$.

Étant donné une BBA $m$, deux mesures peuvent être définis, une *fonction de crédibilité* ($Bel$) et une *fonction de plausibilité* ($Pl$) à l'aide des expressions suivantes :

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

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

$Bel(A)$ peut être interprété comme le degré de soutien total à $A$, tandis que $Pl(A)$ est le degré d'absence de doute sur $A$.

Si son élément focal est $\Omega$, la BBA est appelée *Vide*, et représente la *ignorance* totale.

Deux BBAs $m_1$ et $m_2$ représentant des éléments de preuve indépendants peuvent être combinées par la règle de Dempster définie comme :

$$(m_1 \oplus m_2)(A) = \frac{1}{1-k} \sum_{B \cap C = A} m_1(B) m_2(C) \tag{5}$$

Pour tout $A \subseteq \Omega$, $A \neq \emptyset$, et $(m_1 \oplus m_2)(\emptyset) = 0$. La constante $k$ est appelée le degré de conflit entre les deux BBA et est donnée par :

$$k = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{6}$$

## 3 Méthode proposée

Dans ce travail, une architecture de fusion croisée lidar-caméra basée sur l'apprentissage profond et la formulation évidentielle est proposée. La méthode vise à combiner les forces des deux cadres pour réaliser la segmentation de scènes. Un réseau réduit appelé Lite-CF est combiné à la théorie de l'évidence. À l'origine, Lite-CF est un modèle probabiliste pour la détection des routes [13]. Il produit des distributions de probabilité à partir de logits (les fonctions qui peuvent prendre des valeurs binaires) en utilisant une couche softmax. Pour mieux gérer l'incertitude de la prédiction, l'approche évidentielle est utilisée à la place [1].

### 3.1 Architecture de fusion croisée évidentielle

En conséquence, le Lite-CF évidentiel produit des BBAs plutôt que des probabilités pour représenter l'incertitude de prédiction. L'architecture globale de la Lite-CF évidentielle est donnée dans la Fig. 1. Elle comprend un réseau basé sur un codeur-décodeur, une couche de formulation évidentielle et une unité de décision. Le réseau basé sur un codeur-décodeur comporte deux pipelines de traitement de 18 couches chacun, l'un pour le LiDAR et l'autre pour l'entrée caméra.

Chaque couche d'une modalité est fusionnée avec la couche correspondante de l'autre modalité par une opération de somme pondérée. Les poids de fusion sont entraînables, ce qui permet de fixer la position et l'étendue de la fusion d'être fixées par les données. Une fois que les entrées du LiDAR et de la caméra sont représentées par des BBA dans la couche de formulation évidentielle, une décision peut être prise sur un élément souhaité de l'ensemble $2^\Omega$. Dans le cas considéré de la détection de routes, $2^\Omega$ comprend les éléments "route", "hors route" et "ignorance". De même, pour la tâche étendue de la segmentation sémantique, la décision peut être prise sur n'importe quel élément de $2^\Omega$. Par conséquent, l'approche évidentielle permet une prédiction de classe imprécise, en séparant les données prédites dont le modèle est convaincu des informations erronées.

### 3.2 Formulation évidentielle

La couche de formulation évidentielle utilise en entrée les cartes de caractéristiques (en anglais *features maps*) de la section de décodage. Une fois la résolution complète atteinte dans le décodeur, les cartes de caractéristiques associées (par example L18, dans la Fig. 1) sont utilisées

1. https://github.com/vasigiurgi/evi-cf-deep-learning-based-for-iv

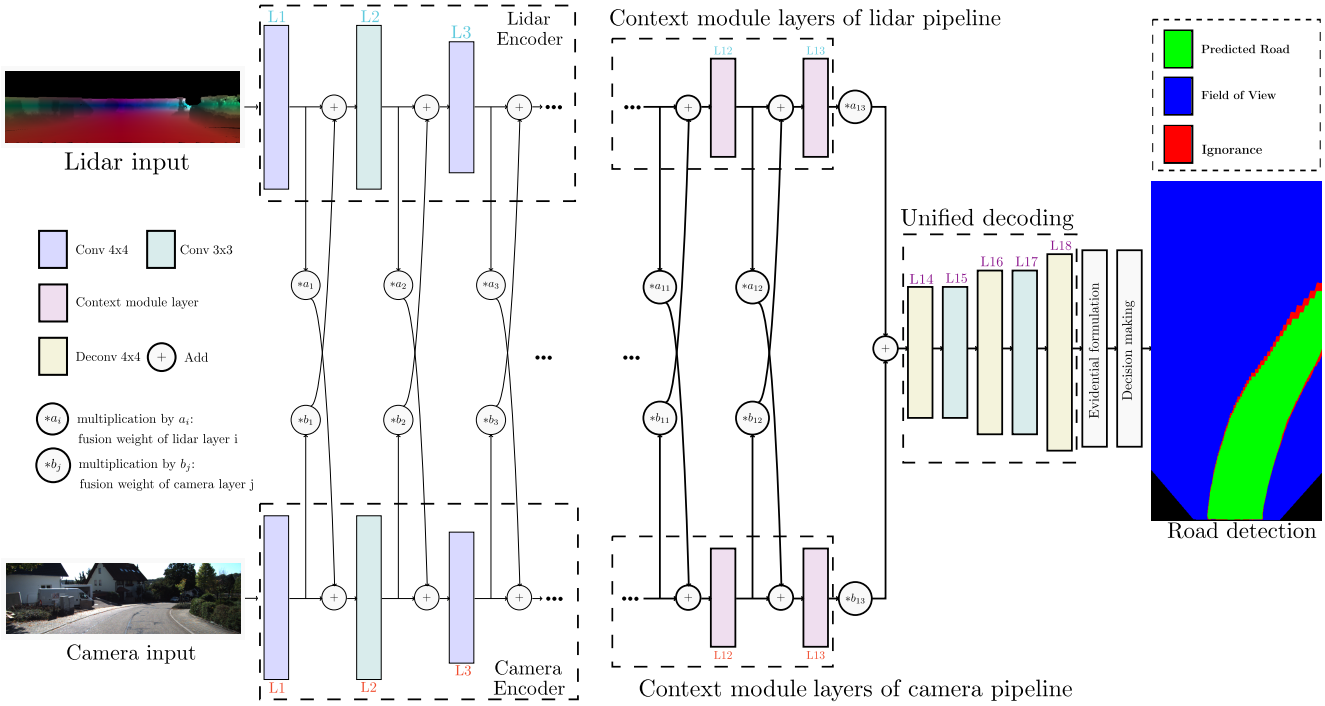



FIGURE 1 – Evidential Lite-CF Architecture. Détection de route en vue d'oiseau (BEV) : Vert : route, Rouge : ignorance

pour construire les BBAs. Le processus est une procédure en trois étapes [16] :

Step 1 : Distance par rapport au prototype : Soit $\boldsymbol{x}$ un vecteur de caractéristiques représentant les caractéristiques d'un pixel à classer éventuellement comme non-road $\omega_1$ ou road $\omega_2$ (par example, le FoD $\Omega = \{\omega_1, \omega_2\}$). La distance euclidienne $d^i$ est déterminée entre $\boldsymbol{x}$ et chaque prototype $\boldsymbol{p^i}$ :

$$d^i = \|\boldsymbol{x} - \boldsymbol{p^i}\| \qquad i = 1, \cdots, n. \tag{7}$$

Step 2 : *Basic belief assignment :* Chaque prototype $\boldsymbol{p^i}$ a un degré d'appartenance $u_j^i$ à chaque classe $\omega_j$, avec une contrainte $u_1^i + u_2^i = 1$. En utilisant l'appartenance à la classe $u_j^i$ et la distance $d^i$, une BBA $m^i$ est construite comme suit :

$$\begin{aligned} m^i(\{\omega_j\}) &= \alpha^i u_j^i \phi^i(d^i), \qquad j = 1, 2 \\ m^i(\Omega) &= 1 - \alpha^i \phi^i(d^i), \end{aligned} \tag{8}$$

où $0 < \alpha^i < 1$ et la fonction $\phi^i$ est définie comme :

$$\phi^i(d^i) = \exp\left(\gamma^i d^i\right), \qquad \gamma^i > 0 \tag{9}$$

Step 3 : Combinaison : Les BBA de l'étape 2 sont combinées en utilisant la règle de Dempster (voir (5)). La BBA combinée résultante représente la preuve permettant de prendre une décision sur la classe de pixel.

Les paramètres associés au prototype $\boldsymbol{p^i}$ (notamment $\alpha^i, u_j^i$, et $\gamma^i$) sont implémentés dans les architectures basées sur l'apprentissage profond et formulés de manière évidentielle comme des poids. Par conséquent, ils sont redéfinis en termes de variables évaluées par des nombres réels, à savoir $\eta^i, \xi^i$ et $\beta_j^i$ :

$$\gamma^i = (\eta^i)^2 \tag{10}$$

$$\alpha^i = \frac{1}{1 + \exp\{-\xi^i\}} \tag{11}$$

$$u_j^i = \frac{(\beta_j^i)^2 + \epsilon}{\sum_{k=1}^{2} (\beta_k^i)^2 + \epsilon} \tag{12}$$

L'équation (12) est légèrement modifiée par rapport à l'expression donnée dans [16]. Un petit nombre positif $\epsilon$ est introduit pour éviter que les valeurs d'appartenance $u_j^i$ ne deviennent nulles. Sinon, les approximations dans une représentation numérique des nombres peuvent faire échouer la règle de Dempster en raison du *conflit total*.

## 3.3 Prise de décision

Une fois que les BBA représentant les éléments de preuve dans les pixels correspondants sont évalués, la tâche finale consiste à décider des classes des pixels. Dans ce travail, une décision basée sur la distance d'intervalle de croyance est utilisée [17]. Pour réduire le coût de calcul associé, la règle de décision est simplifiée pour le cas de la détection des routes :

Case i) La décision est limitée aux singletons : Les éléments de décision possibles sont $\omega_1$ (pas de

route) et $\omega_2$ (route). Dans ce cas, la règle de décision devient :

$$\hat{X} = Arg \max_{X \in \{\omega_1, \omega_2\}} m(X). \quad (13)$$

Case ii) La décision n'est pas contrainte : Il peut être intéressant de permettre l'affectation de pixels ambigus à des classes imprécises comme $\Omega$. Cela peut réduire l'erreur de classification en diminuant la décision qui a plus de nature arbitraire. Le tableau 1 donne l'ensemble des règles lorsque la décision est sans contrainte.

TABLE 1 – La règle de décision sans contrainte [cardinalité du FoD = 2]

| |
|---|
| **If** $m(\{\omega_1\}) > m(\{\omega_2\})$ **and** $m(\{\omega_1\}) > 0.5m(\{\omega_2\}) + 0.5$ **then** |
| Decide $\omega_1$ |
| **Else If** $m(\{\omega_1\}) < m(\{\omega_2\})$ **and** $m(\{\omega_1\}) < 2m(\{\omega_2\}) - 1$ **then** |
| Decide $\omega_2$ |
| **Else** |
| Decide $\Omega$ |

# 4 Résultats expérimentaux

## 4.1 Dataset KITTI et Pré-traitement du LiDAR

Le dataset KITTI sur la tâche de détection des routes est utilisé dans ce travail. Ce ensemble de données contient 289 images pour l'entraînement. Ces images sont classées en trois types de sous-catégories de routes. L'ensemble de données contient des images de caméra et un scan LiDAR 3D. Un point LiDAR 3D $x$ est projeté vers un point $y$ dans le plan de la caméra selon les matrices de projection KITTI $P$, de rectification $R$ et de translation $T$ :

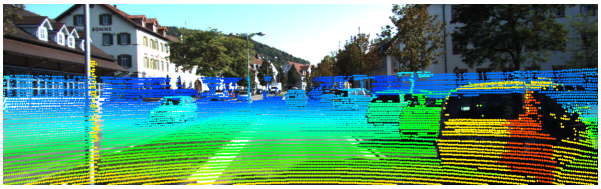
$$y = P\,R\,T\,x \quad (14)$$

Un sur-échantillonnage est utilisé pour produire une carte de profondeur dense, comme le montre la Fig. 2. Le sur-échantillonnage est effectué selon la technique décrite dans [18].

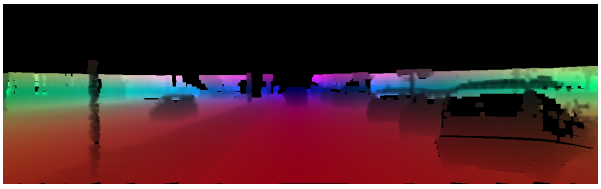
## 4.2 Résultats de la détection des routes

Le modèle évidentiel est entraîné et évalué sur la tâche de détection de routes. Lors de l'entraînement, les masques de la vérité-terrain de KITTI (notamment hors-route = 0, route = 1) sont codés en utilisant l'approche évidentielle : La hors-route est étiquetée comme $[m(\omega_1) = 1, m(\omega_2) = 0, m(\Omega) = 0]$ et la route comme $[m(\omega_1) = 0, m(\omega_2) = 1, m(\Omega) = 0]$. L'évaluation se fait selon la validation croisée stratifiée à 10 reprises.

Le CF de base et sa réduction Lite-CF sont des modèles probabilistes et sont évalués classiquement. Cependant, le Lite-CF évidentiel peut être évalué en utilisant la règle de décision contrainte ou non contrainte (voir la section 3.3). Lorsque la décision est contrainte, elle est nommée Lite-CF-Evi1, sinon Lite-CF-Evi2. L'évaluation utilise les métriques KITTI : score F1 maximum $MaxF$, précision $PRE$, et rappel $REC$. En outre, le taux d'erreur $ER$ et le nombre d'images par seconde *FPS* sont utilisés comme métriques communes à tous les modèles. Le tableau 2 donne l'évaluation des performances en termes de moyenne et d'écart type parmi les 10 répétitions.



(a) projection LiDAR



(b) LiDAR up-sampling

FIGURE 2 – Pré-traitement LiDAR

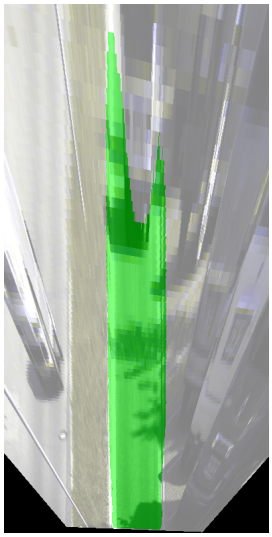
Comme on peut le voir dans le tableau, Lite-CF-Evi1 a le $MaxF$ le plus élevé tout en ayant une réduction d'environ 15,7% des paramètres du modèle par rapport à la ligne de base. L'augmentation de la valeur $MaxF$ et le faible écart-type associé peuvent être attribués à la représentation et à la fusion des éléments de preuve provenant des entrées capteurs et de la règle de décision associée utilisés. En outre, le modèle évidentiel a un temps d'exécution plus rapide que le modèle de base. Ceci est dû au fait que sa base est sur une architecture réduite. De plus, certaines couches, comme les logits et les softmax, sont supprimées dans sa formulation.

En outre, la simplification de la règle de décision a également contribué à l'amélioration du temps d'exécution. Par conséquent, un modèle évidentiel plus petit en taille, plus précis et plus rapide que le modèle de base est réalisé. Une réduction supplémentaire des erreurs au prix de la vitesse est également obtenue en utilisant la règle de décision sans contrainte : Lite-CF-Evi2 donne une réduction d'environ 27% de l'ER à un coût d'environ 12% de perte de FPS par rapport à Lite-CF-Evi1. Le modèle évidentiel, en plus des améliorations de performance obtenues, donne au décideur des options basées sur différentes conditions : il peut donner une prédiction binaire comme pas-route et route ou une prédiction multi-classe qui inclut l'ignorance.

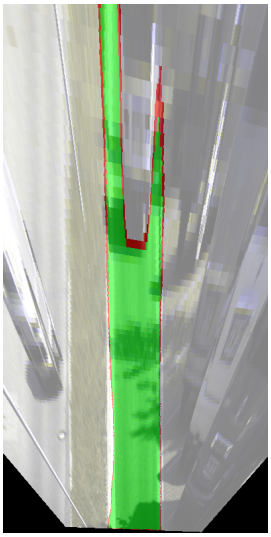
Une comparaison visuelle entre le Lite-CF probabiliste et sa contrepartie évidentielle Lite-CF-Evi2 est présentée dans la Fig. 3. Le Lite-CF-Evi2 possède une classe de prédiction supplémentaire (c'est-à-dire l'ignorance $\Omega$). Comme on peut le voir sur la figure (voir Fig. 3c), la zone

de la route autour du véhicule de tête et du bord de la route est classée comme ignorance dans la Lite-CF-Evi2. Cette classification est appropriée car cette zone est souvent représentée par des pixels ambigus. De plus, les points éloignés sont également classés dans la catégorie ignorance. Ceci est également convaincant car les points extrêmes ne sont pas suffisamment représentés par la caméra ou le LiDAR. En outre, la prédiction évidentielle donne un tracé routier naturel et lisse. De plus, l'arrière de la voiture de tête est mieux segmenté que dans l'approche probabiliste (voir Fig. 3b). Par conséquent, ces aspects augmentent la performance globale de l'approche évidentielle.

(a) Vue en perspective



(b) Probabiliste : Lite-CF



(c) Evidentiel : Lite-CF-Evi2

FIGURE 3 – La vue en perspective et la projection en vue d'oiseau sur la détection des routes.

### 4.3 De la détection des routes à la segmentation sémantique

Ensuite, le modèle de fusion croisée évidentielle est mis au défi pour la tâche de segmentation sémantique. L'architecture est étendue aux images multi-classes du benchmark sémantique KITTI au niveau du pixel. Ce ensemble de données ciblé fournit 200 images de caméra analogiques au benchmark KITTI Stereo and Flow 2015. Comme l'ensemble de données dédié à la segmentation sémantique ne fournit pas de nuages de points LiDAR, la correspondance est faite entre les images de caméra de segmentation sémantique et le LiDAR brut fourni par KITTI [19]. Ainsi, 127 des 200 images caméra sont identifiées, ainsi que leurs trames LiDAR correspondantes. Ces dernières sont projetées, sur-échantillonnées en images de profondeur denses [13], [18] et finalement fusionnées au modèle Lite-CF-Evi avec les images de caméra, comme dans le cas de la détection de route.

Compte tenu de sa faible complexité, mais de ses bonnes performances dans les scénarios de segmentation routière, l'architecture Lite-CF-Evi est étendue à des tâches plus complexes, incluant des classes multiples, comme la segmentation multi-classes (segmentation sémantique). Les classes sont simplifiées pour 3 objets : "route", "voiture", et "arrière-plan". Dans ce cas, l'arrière-plan représente tout ce qui est autre que les routes et les voitures. L'ensemble de données est divisé en 114 images pour l'entraînement et 13 images pour la validation.

Lite-CF-Evi obtient des résultats intéressants quelle que soit la complexité de l'application. Pour évaluer les performances, le modèle est évalué avec la métrique d'intersection-sur-union $IoU$, conformément au PASCAL VOC [20] :

$$IoU = \frac{TP}{TP + FP + FN} \tag{15}$$

où TP, FP et FN sont respectivement les vrais positifs, les faux positifs et les faux négatifs.

L'architecture d'apprentissage profond évidentiel est évaluée sur l'ensemble de validation pour les 3 classes. Le modèle obtient une moyenne de $IoU = 93.17$ pour le jeu de validation donné. Les valeurs IoU affichent des performances élevées. Cela peut s'expliquer par le nombre de classes qui est simplifié (il n'est donc pas exponentiellement plus complexe que la détection de routes) et leurs valeurs reflètent également la surface de l'objet, pixel par pixel (les objets minuscules qui rendraient la prédiction plus difficile ne sont pas pris en compte).

Par conséquent, l'évolutivité de la théorie de Dempster-Shafer est mise en évidence : le modèle est flexible, de sorte qu'en changeant la valeur des éléments de décision, le nombre d'éléments traités de l'ensemble de puissance peut être facilement modifié également. Il influence directement le nombre de classes prédites. À cet endroit, les classes sont représentées par des eleménts, des affectations de pixels à des éléments de puissance non nulle. La prise de décision peut être adaptée pour évoluer d'un nombre fixe de classes (égal au nombre de singletons) au nombre maximum des elements, $2^{\Omega}-1$, où $\Omega$ est le cadre de discernement, comme présenté dans la section 2.

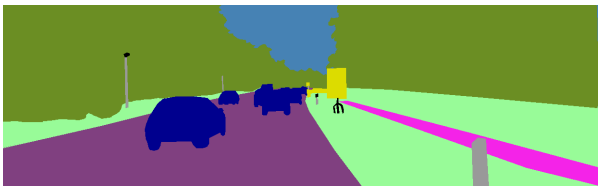
En particulier, dans cette tâche de segmentation sémantique, 3 singletons sont initialement considérés : la route, la voiture et l'arrière-plan. De plus, la formulation évidentielle introduit la quatrième classe, à savoir "l'ignorance" pour traiter les prédictions incertaines. Dans les illustrations suivantes, il est présenté un ensemble de prédictions évaluées avec le modèle Lite-CF-Evi. La Fig. 4a représente l'image originale provenant de la caméra. Parallèlement, l'architecture reçoit en entrée une image LiDAR

TABLE 2 – Comparaison des performances des modèles

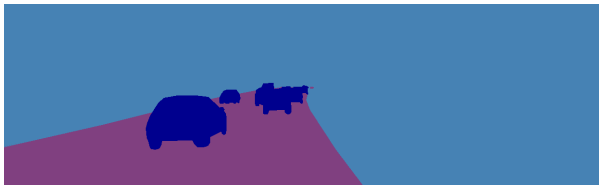
| Model arch. | # model param. | MaxF | PRE | REC | ER | FPS |
|---|---|---|---|---|---|---|
| Baseline (CF)[9, 13] | 3,246,830 | $96.25 \pm 0.71$ | $96.46 \pm 0.66$ | $96.05 \pm 1.06$ | $1.34 \pm 0.26$ | 27 |
| Lite-CF [13] | 2,737,213 | $95.50 \pm 0.52$ | $95.57 \pm 0.69$ | $95.45 \pm 0.74$ | $1.61 \pm 0.21$ | **35** |
| Lite-CF-Evi1 | **2,737,066** | $\mathbf{96.91 \pm 0.36}$ | $\mathbf{96.74 \pm 0.56}$ | $\mathbf{97.09 \pm 0.71}$ | $1.11 \pm 0.14$ | 33 |
| Lite-CF-Evi2 | **2,737,066** | - | - | - | $\mathbf{0.81 \pm 0.13}$ | 29 |

synchronisée, comme présenté dans la Fig. 2b. En continu, la deuxième image, Fig. 4b représente la vérité de terrain RVB complète, avec des classes annotées comme dans le repère de segmentation sémantique pixellisée KITTI et les paysages urbains. La troisième image (Fig. 4c) illustre la vérité terrain simplifiée utilisée dans ce travail, avec 3 classes : la route (magenta), la voiture (bleu foncé) et l'arrière-plan (bleu, correspondant à la classe ciel de l'annotation originale). La classe voiture, à son tour, est composée d'étiquettes de voitures, de camions et de bus. La classe arrière-plan encapsule toutes les classes, à l'exception de la route et des voitures. Enfin, la dernière image, Fig. 4d représente l'image prédite avec Lite-CF-Evi. On peut observer que les classes route, voiture et arrière-plan sont prédites avec précision, quelle que soit la complexité de la tâche. De plus, la classe supplémentaire, l'ignorance (illustrée en blanc) exprime bien les pixels appartenant à des prédictions imprécises, et elle évite d'allouer les pixels ambigus à une classe inappropriée comme cela pourrait arriver en utilisant une approche probabiliste. L'ignorance est prépondérante dans les régions où se trouvent les frontières entre les classes, où le modèle est plus vulnérable aux prédictions erronées. Comme dans le cas de la détection des routes, dans la prédiction multi-classes, les points éloignés représentent des informations manquantes. Cela pourrait symboliser le fait que le modèle manque de confiance dans la classification des pixels correspondant aux objets éloignés en raison de données incertaines. Par conséquent, ces pixels sont étiquetés comme étant de l'ignorance, ce qui donne une meilleure compréhension et prouve que le raisonnement évidentiel fonctionne bien dans la gestion des incertitudes.

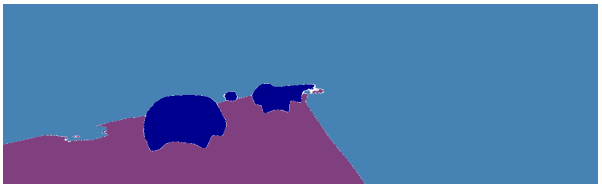
(a) Image de l'appareil photo issue de la segmentation sémantique KITTI



(b) Vérité terrain originale : RVB sémantique



(c) Vérité terrain simplifiée, 3 classes : route, voiture, arrière-plan



(d) Prédiction avec Lite-CF-Evi : route, voiture, fond, ignorance (blanc)

FIGURE 4 – Résultats de la segmentation sémantique

## 5 Conclusion

Dans cet travail, une fusion caméra-lidar est proposée en utilisant une architecture d'apprentissage profond combinée aux fonctions de croyances pour la perception de l'environnement dans les véhicules intelligents. Cette combinaison permet de mieux représenter l'incertitude des prédictions du réseau. Dans la tâche de détection des routes, l'approche évidentielle surpasse le modèle probabiliste. Elle donne un MaxF élevé (faible ER) tout en considérant la limitation de calcul dans les VI. L'introduction de l'ignorance comme élément de décision améliore encore les performances. Ainsi, les points éloignés et les caractéristiques ambiguës peuvent être classés dans la catégorie de l'ignorance. L'approche est étendue à des tâches plus complexes, telles que la segmentation sémantique. On observe que les modèles d'apprentissage profond basés sur les fonctions de croyance peuvent facilement être étendus à des applications de reconnaissance de routes à la segmentation multi-classes, et que la formulation évidentielle est générique. Pour obtenir des améliorations cohérentes pour les prédictions de segmentation sémantique, une analyse

plus approfondie du modèle est envisagée. Le modèle Lite CF-Evi vise à être mis à jour en ce qui concerne la complexité et le coût de calcul, en gardant à l'esprit l'objectif d'une mise en œuvre en temps réel. Une combinaison de plusieurs architectures de fusion et de règles de décision alternatives est étudiée.

# 6 Remerciements

Les auteurs tiennent à remercier le projet de l'Agence Nationale de la Recherche (projet EviDeep ANR JCJC), l'Université de Haute-Alsace, le Ministère de l'Education Ethiopien (Ambassade de France en Ethiopie et l'Union Africaine dans le cadre du programme Ethio-Français de bourses de doctorat en ingénierie) et la Fondation Pierre-et-Jeanne Spiegel pour le soutien apporté à la réalisation de ce projet.

# Références

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

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

[4] H. Laghmara, M.-T. Boudali, T. Laurain, J. Ledy, R. Orjuela, J.-P. Lauffenburger, and M. Basset, "Obstacle avoidance, path planning and control for autonomous vehicles," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 529–534.

[5] P. Xu, F. Davoine, and T. Denoeux, "Evidential combination of pedestrian detectors," in *British Machine Vision Conference*, 2014, pp. 1–14.

[6] P. Xu, F. Davoine, J.-B. Bordes, H. Zhao, and T. Denœux, "Multimodal information fusion for urban scene understanding," *Machine Vision and Applications*, vol. 27, no. 3, pp. 331–349, 2016.

[7] E. Capellier, F. Davoine, V. Cherfaoui, and Y. Li, "Evidential deep learning for arbitrary lidar object classification in the context of autonomous driving," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1304–1311.

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

[9] 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.

[10] L. Caltagirone, S. Scheidegger, L. Svensson, and M. Wahde, "Fast lidar-based road detection using fully convolutional neural networks," in *2017 ieee intelligent vehicles symposium (iv)*. IEEE, 2017, pp. 1019–1024.

[11] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, "Lidar-based driving path generation using fully convolutional neural networks," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2017, pp. 1–6.

[12] D.-V. Giurgi, T. Josso-Laurain, M. Devanne, and J.-P. Lauffenburger, "Real-time road detection implementation of unet architecture for autonomous driving," in *IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 2022, pp. 1–5.

[13] 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 International Conference on Computers and Automation (CompAuto)*. IEEE, 2022, p. to be given.

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

[15] T. Denœux, D. Dubois, and H. Prade, "Representations of uncertainty in ai : beyond probability and possibility," in *A guided tour of artificial intelligence research*. Springer, 2020, pp. 119–150.

[16] 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.

[17] 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.

[18] C. Premebida, J. Carreira, J. Batista, and U. Nunes, "Pedestrian detection combining rgb and dense lidar data," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 4112–4117.

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

[20] M. Everingham, S. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge : A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.