Reliability-Aware Fusion for Semantic Segmentation under Sensor Degradation and Failures

Abdelhak Benamirouche[1,*], Lucas Deregnaucourt[2,*], Mihreteab Negash Geletu[1]
Hind Laghmara[2], Remi Boutteau[2], Jean-Philippe Lauffenburger[1]

[1] IRIMAS-UR7499, Université de Haute-Alsace, Mulhouse, France.
[2] INSA Rouen Normandie, Univ Rouen Normandie, Université Le Havre Normandie, Normandie Univ, LITIS UR 4108, F-76000 Rouen, France.
[*] Equal contribution.

Résumé

La segmentation sémantique dans des scénarios de conduite réels est particulièrement difficile en raison de la dégradation des capteurs, des pannes et des conditions environnementales changeantes. Bien que la fusion multimodale soit une solution couramment utilisée, de nombreuses approches existantes traitent toutes les modalités de manière équivalente, en ignorant leur fiabilité variable selon les classes sémantiques et les conditions. Dans cet article, nous présentons ReCoLaF (Reliability-aware Conflict-guided Late Fusion), un nouveau cadre de fusion profonde pour la segmentation sémantique multimodale en présence d'incertitude. ReCoLaF ajuste de manière adaptative la contribution de chaque modalité de capteur grâce à une stratégie de pondération en deux étapes : un module de fiabilité appris, spécifique aux classes, qui estime la pertinence de chaque modalité pour différentes classes sémantiques, ainsi qu'un ajustement basé sur le conflit qui mesure les incohérences locales entre modalités au niveau du pixel. La fusion est formulée dans le cadre de la théorie de l'évidence de Dempster-Shafer, offrant une approche mathématiquement fondée pour gérer l'incertitude et produire des prédictions robustes. Nous évaluons ReCoLaF sur les jeux de données DeLiVER (synthétique) et MUSES (réel) dans diverses conditions météorologiques et configurations de capteurs dégradés. ReCoLaF obtient systématiquement de meilleures performances moyennes en présence de défaillances de capteurs, mettant en évidence l'intérêt de modéliser conjointement la fiabilité sémantique et l'accord inter-modalités pour une fusion robuste dans des scénarios de conduite complexes.

Mots-clés

Segmentation sémantique, Fusion multimodale, Dégradation des capteurs, Théorie de Dempster-Shafer, Conduite autonome.

Abstract

Semantic segmentation in real-world driving scenarios is particularly challenging due to sensor degradation, failures, and changing environmental conditions. While multimodal fusion is a common solution, many existing approaches treat all modalities equally, ignoring their varying reliability across semantic classes and conditions. In this paper, we present ReCoLaF (Reliability-aware Conflict-guided Late Fusion), a novel deep fusion framework for multimodal semantic segmentation under uncertainty. ReCoLaF adaptively adjusts the contribution of each sensor modality through a two-stage weighting strategy: a learned class-wise reliability module that estimates how relevant each modality is for different semantic classes, and a conflict-based adjustment that measures local inconsistencies between modalities at the pixel level. The fusion is formulated within the Dempster-Shafer theory of evidence, providing a mathematically grounded approach to handle uncertainty and make robust predictions. We evaluate ReCoLaF on the DeLiVER (synthetic) and MUSES (real-world) datasets under diverse weather conditions and degraded sensor configurations. ReCoLaF consistently achieves higher average performance under sensor failures, highlighting the benefit of jointly modeling semantic reliability and inter-modality agreement for robust fusion in complex driving scenarios.

Keywords

Semantic segmentation, Multimodal fusion, Sensor degradation, Dempster-Shafer theory, Autonomous driving.

1 Introduction

Semantic segmentation is a critical component of environment perception in autonomous driving systems. However, performance can be significantly degraded in real-world scenarios due to both external and internal factors. Externally, the driving scene is often unstructured, with occlusions, object truncations, and variable weather conditions such as rain, fog, or snow. Illumination changes from day to night or reflections further complicate scene understanding. Internally, perception sensors have inherent limitations: cameras provide rich color and texture but

are sensitive to lighting; LiDAR and radar are robust to illumination but have lower resolution; event cameras handle motion and lighting variations well but lack color information. Moreover, sensor failures or degradation can occur during deployment, making the need for robust perception ever more pressing.

To overcome these challenges, multimodal sensor fusion is widely used to integrate complementary information and to improve the robustness of perception systems. Recent deep learning-based fusion architectures show high results. They combine modalities at various stages using operations such as concatenation, addition, ensembles or mixtures of experts [8]. These methods have been applied to object detection, road segmentation, and semantic scene understanding [1, 21]. More recently, evidential deep learning has been introduced to model uncertainty explicitly in multimodal fusion [20, 10]. Based on evidence theory, these models provide extended mechanisms for uncertainty handling and more informed fusion strategies.

In real-world driving scenes, different modalities are not equally informative for all object classes. For instance, LiDAR may perform better for detecting structures, while RGB excels at recognizing signs or road markings. Moreover, in degraded settings—such as fog or partial sensor failure—some modalities become unreliable or even detrimental. To address this, robust fusion should dynamically adjust modality contributions based on their relevance and consistency. In this work, we propose an evidential deep fusion framework for semantic segmentation that adaptively fuses multiple modalities by jointly modeling both modality-specific class-level reliability and inter-modality conflict.

The organization of this paper is as follows. Section 2 reviews related work in semantic and multimodal perception, with a focus on recent approaches to evidential fusion and sensor-aware modeling. Section 3 introduces the fundamentals of evidence theory as the underlying framework for our evidential reasoning. Section 4 presents our proposed architecture, ReCoLaF, detailing its four main components: evidential encoder-decoders, sensor reliability estimation, conflict-based adjustment, and evidence fusion. Section 5 describes the experimental setup, datasets, and implementation details, followed by a thorough evaluation of our method across various sensor combinations and conditions. We conclude the paper in Section 6 with a summary of contributions and directions for future research.

2 Related work

Robust and accurate scene understanding can be critical under diverse and often challenging environmental conditions. Single-modality sensors, such as RGB cameras, LiDAR, and radar, each offer unique advantages —color and texture, precise depth, and weather resilience, respectively— but also exhibit limitations when faced with low visibility, adverse weather, or sensor degradation. To address these challenges, **multimodal perception** integrates complementary information from multiple sensors, enabling perception systems to compensate for the weaknesses of individual modalities and improving performance in tasks such as detection, segmentation, localization, and navigation. Early efforts focused on enhancing LiDAR-based 3D detection with RGB data, supported by benchmark datasets like KITTI [9] and nuScenes [4]. However, these lacked coverage of adverse conditions. Fusion strategies have evolved from fixed dual-modality combinations to flexible RGB-X fusion frameworks capable of handling arbitrary modality sets. Recent methods such as CMNeXt [23] employ modular attention-based architectures, while approaches like StitchFusion [13] explore large-scale pre-trained backbones and modality-specific encoding.

While these works enhance perception in challenging environments, they typically fuse sensor inputs uniformly and do not explicitly model the reliability or contextual relevance of each modality. To address this, Huang et al. [11] proposed a fusion framework under the Dempster-Shafer theory that introduces contextual discounting based on class-specific reliability estimation for each modality. This approach adjusts the influence of each sensor depending on its expected semantic reliability, allowing more informed evidence fusion. Other approaches have explored how inter-sensor disagreement can guide adaptive fusion. Deregnaucourt et al. [5] introduced ECoLaF, a conflict-guided fusion framework that discounts modality contributions based on local disagreement with other modalities. ECoLaF demonstrated strong robustness in degraded conditions by dynamically adjusting trust in each sensor at the pixel level. In parallel, CAFuser [3] explored environmental condition-aware fusion mechanisms, enabling models to adapt to external factors such as fog, rain, or night-time illumination, rather than relying solely on the sensor signal itself.

In this work, we propose a novel and robust evidential fusion framework that unifies the strengths of both contextual and conflict-guided discounting. Specifically, we estimate sensor reliability per class to discount mass functions semantically, and subsequently apply a conflict-guided discounting step to account for disagreement across modalities at the spatial level. This two-stage trust modeling mechanism allows our model to reason both about the expected utility of each sensor and its consistency with others, enabling robust and adaptive multimodal segmentation under uncertain or degraded conditions.

3 Evidence Theory Basics

Evidence theory, or Dempster-Shafer theory, is a formalism for representing, reasoning and making decision under uncertainty [19]. Two major uncertainty types can be distinguished: it is known to be aleatory when due to process randomness and can not be reduced. When uncertainty rises from a lack of knowledge, it is epistemic. This form of uncertainty can be limited by acquiring additional information.

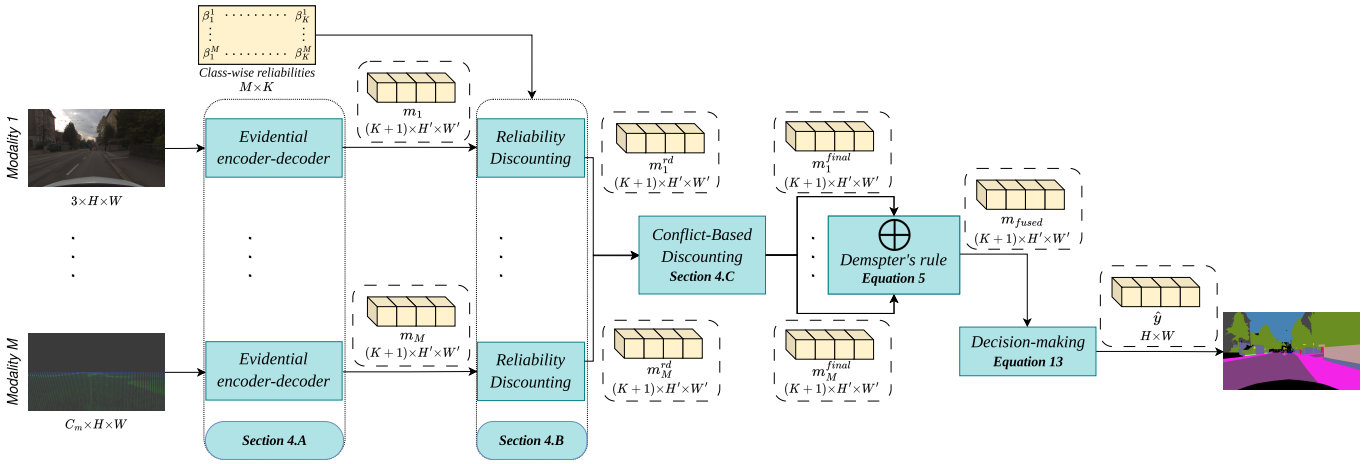

Figure 1: **ReCoLaF architecture.** Each modality is associated to an independent evidential encoder-decoder, which outputs mass functions. The mass functions of each modality are first discounted on the basis of their reliability and then on the basis of their respective conflict. The discounted mass functions are then fused and converted into probabilities to make a decision.

3.1 General definitions

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

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

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

where 2^Ω is the power set of Ω defined as follows:

$$2^\Omega = \{\emptyset, \{\omega_1\}, \ldots, \{\omega_n\}, \{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}, \ldots, \Omega\} \tag{3}$$

For clarity and coherence purpose, m will be called a mass function in the following sections. For any subset $A \in 2^\Omega$, $m(A)$ is bounded between 0 and 1. 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.

3.2 Mass function discounting

A source of evidence may not be reliable or its associated support can be inaccurate. In this situation discounting the support given by the mass function is relevant [19]. Consider $1 - \alpha$, the discounting factor, i. e. α the degree of trust in the evidence with $0 < \alpha < 1$. The discounted mass function ${}^\alpha m(A)$ is given as:

$$\begin{aligned} {}^\alpha m(A) &= \alpha \cdot m(A) \quad \forall A \subset \Omega \\ {}^\alpha m(\Omega) &= (1 - \alpha) + \alpha \cdot m(\Omega) \end{aligned} \tag{4}$$

Discounting can be used to lower the effect of sources which are not fully trusted before evidence combination (see Section 3.3).

3.3 Evidence Combination and Conflict Management

Two mass functions m_1 and m_2 representing independent pieces of evidence (e.g., predictions from two different sensors) on a common frame Ω can be combined by Dempster-Shafer's rule (DS) [19] defined as:

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

for all $A \in 2^\Omega$, $A \neq \emptyset$, and $(m_1 \oplus m_2)(\emptyset) = 0$. The DS rule is commutative and associative. The constant k is called the conflict degree between the mass functions and is given as:

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

If k=0, the mass functions m_1 and m_2 are 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) and their combination through the DS rule is impossible.

3.4 Decision in evidence theory

Once the sources of evidence are combined, a probabilistic decision rule is required to select the final class $\omega_k \in \Omega$. In this work, we adopt the DSmP transformation [6], which redistributes the uncertainty toward singletons while minimizing the entropy of the resulting probability distribution.
The DSmP transformation for a given singleton ω_k is defined as:

$$\text{DSmP}_\varepsilon(\omega_k) = \sum_{A \in 2^\Omega} m(A) \frac{\sum_{a \in \{\omega_k \cap A\}} m(a) + \varepsilon \cdot |\omega_k \cap A|}{\sum_{a \in A} m(a) + \varepsilon \cdot |A|} \tag{7}$$

The parameter $\varepsilon > 0$ moderates the impact of the focal elements cardinality in the uncertainty redistribution.

4 Reliability-Aware Fusion Framework

Figure 1 illustrates the overall architecture of the proposed adaptive evidential fusion framework. The model is composed of four main stages: modality-specific evidential encoder-decoders, sensor reliability estimation, conflict-based discounting, and evidential fusion through Dempster-Shafer theory. We detail each component of the architecture in the following subsections.

4.1 Evidential Encoder-Decoder (per modality)

Let $\mathcal{M}=\{1,\ldots,M\}$ be the set of modalities and $\Omega=\{\omega_1,\ldots,\omega_K\}$ the FoD for a K-class segmentation task. For each modality $m \in \mathcal{M}$, an encoder-decoder produces a mass function map $m_m^{\text{raw}}(i,j) \in \mathbb{R}^{K+1}$ for each pixel (i,j) by applying a softmax activation function to ensure that the mass functions sum to one.

Following prior work in evidential deep learning [20, 5], we restrict the set of focal elements to only the singletons and the full frame Ω. That is, for each class $\omega_k \in \Omega$, a mass value is assigned to $\{\omega_k\}$, and the remaining mass is assigned to Ω, which captures total uncertainty. This simplifies the mass function representation while retaining the expressiveness needed for uncertainty modeling.

4.2 Class-wise Sensor Reliability

We assign to each modality a vector of class-specific reliability scores:

$$\boldsymbol{\beta}_m = (\beta_1^m, \beta_2^m, \ldots, \beta_K^m) \in [0,1]^K \tag{8}$$

Each β_k^m quantifies how reliable modality m is for predicting class ω_k. These scores are used to discount each of the raw mass functions per class, resulting in the discounted mass function m_m^{rd}:

$$m_m^{\text{rd}}(\{\omega_k\}) = \beta_k^m \cdot m_m^{\text{raw}}(\{\omega_k\}) \tag{9}$$

$$m_m^{\text{rd}}(\Omega) = 1 - \sum_{k=1}^{K} m_m^{\text{rd}}(\{\omega_k\}) \tag{10}$$

The reliability scores for each modality $\boldsymbol{\beta}_m$ are implemented as trainable parameters and are learned jointly with the rest of the network during training. Importantly, during inference time, they are fixed, not input-dependent. This design reflects a global semantic prior that captures the average reliability of each modality for each class.

4.3 Conflict-based Adjustment

To account for sensor disagreements, we apply a conflict-based discounting mechanism. Following the ECoLaF framework [5], we measure the level of disagreement between modalities using the Jousselme distance [12] and derive a conflict score Conf_m for each modality. This score is then converted into a conflict-based reliability score α_m [18].

The final mass function m_m^{final} is given by the discounting formula:

$$\begin{cases} m_m^{\text{final}}(\omega_k) & = \alpha_m \cdot m_m^{\text{rd}}(\omega_k), \qquad \omega_k \in \Omega \\ m_m^{\text{final}}(\Omega) & = 1 - \alpha_m + \alpha_m \cdot m_m^{\text{rd}}(\Omega) \end{cases} \tag{11}$$

This step adjusts the influence of each sensor locally based on how much it agrees with others.

4.4 Multimodal Fusion and Class Definition

The refined mass functions from all modalities are fused using Dempster's rule $\oplus$ (Eq. 5):

$$m_{\text{fused}}(A) = \left(\bigoplus_{m=1}^{M} m_m^{\text{final}} \right)(A) \tag{12}$$

For decision making, according to (7), we choose the class with highest DSmP as the final prediction:

$$\hat{y} = \arg\max_k \text{DSmP}_\varepsilon(\omega_k) \tag{13}$$

Following the recommendations in [6], we choose $\varepsilon = 0.001$ to obtain a probability function with an entropy as low as possible while avoiding numerical instabilities.

In our architecture of combining the model-based approach of Dempster-Shafer theory with deep learning, the reliability weighting is positioned ahead of the conflict discounting (see Fig. 1). In this arrangement, semantically relevant mass functions are obtained first by class-based rectification (i.e., class-specific reliability value β_k^m, (Eq. 8)). Then, they are discounted based on their inter-class disagreement. If the order were the opposite, the conflict among sources could be calculated based on less refined mass functions. Therefore, the proposed order of computations is chosen for maintaining information quality.

5 Experiments

5.1 Datasets and Implementation Details

A quantitative validation has been performed on two multimodal datasets designed for robust semantic segmentation under challenging driving conditions: the **DeLiVER** [24] synthetic dataset, and **MUSES** [2], a non-synthetic real-world dataset. Both datasets provide diverse driving scenarios and multimodal sensor data.

DeLiVER comprises paired images from four sensor modalities—RGB, Depth, Event, and LiDAR—captured under diverse weather conditions and simulated sensor failure scenarios, including motion blur, over-exposure, under-exposure, LiDAR jitter, and low-resolution Event data. It is designed to study the semantic segmentation of road scenes across 25 classes. The dataset includes 3983, 2005, and 1897 front-view image pairs for training, validation, and testing, respectively.

MUSES is a dataset dedicated to the analysis of real-world road scenes under adverse weather conditions, specifically fog, rain, and snow. It comprises 1500, 250, and 750 paired images across four modalities—RGB, Event, LiDAR, and RADAR—for training, validation, and testing,

RGB	Depth	Event	LiDAR	CMNeXt [24]	CAFuser [3]	ECoLaF [5]	ReCoLaF
✓				20.62	24.69	31.44	**35.62**
	✓			40.29	**43.52**	38.77	42.06
		✓		2.82	1.74	1.87	**3.77**
			✓	2.76	1.44	2.40	**3.13**
✓	✓			52.96	**54.05**	49.23	49.89
✓		✓		20.37	26.29	31.44	**35.69**
✓			✓	20.79	26.04	31.72	**35.55**
	✓	✓		40.46	**43.95**	38.77	42.48
	✓		✓	40.29	**43.52**	38.86	42.05
		✓	✓	2.81	1.58	2.03	**4.14**
✓	✓	✓		53.11	**54.44**	49.23	49.90
✓	✓		✓	52.88	**53.93**	49.25	49.90
✓		✓	✓	20.54	27.37	31.72	**35.66**
	✓	✓	✓	40.39	**43.78**	38.86	42.58
✓	✓	✓	✓	53.01	**53.87**	49.25	49.90
mean				30.94	33.35	32.35	**34.82**

(a) DeLiVER

RGB	Event	LiDAR	Radar	CMNeXt [24]	CAFuser [3]	ECoLaF [5]	ReCoLaF
✓				49.82	63.92	65.02	**66.56**
	✓			2.65	**4.51**	3.15	3.99
		✓		2.65	16.79	19.49	**33.90**
			✓	7.56	**8.34**	3.64	4.23
✓	✓			52.55	65.33	65.02	**66.56**
✓		✓		66.90	**68.57**	66.43	68.29
✓			✓	61.66	65.10	65.02	**66.56**
	✓	✓		2.62	16.83	19.49	**33.90**
	✓		✓	7.71	**8.39**	3.64	4.23
		✓	✓	9.94	15.18	20.20	**34.04**
✓	✓	✓		66.64	**68.76**	66.43	68.29
✓	✓		✓	62.09	65.26	65.02	**66.56**
✓		✓	✓	71.06	**72.88**	66.42	68.29
	✓	✓	✓	10.82	17.59	20.20	**34.04**
✓	✓	✓	✓	71.06	**72.88**	66.42	68.29
mean				36.38	42.02	41.04	**45.85**

(b) MUSES

Table 1: Performances comparison of using different modalities in mIoU(%). Each row represents a test-time inference scenario where a subset of modalities is available (✓) and the others are disabled (i.e., replaced with zero-filled tensors to simulate sensor failure). ✓ indicates the available modalities at test time, while others are disabled (zero-filled) to simulate sensor failures. Bold values represent the best performances to the nearest rounding.

respectively. As ground truth annotations are not available for the original test set, we re-split the training set into a new training and validation subset, and repurpose the original validation set as the new test set. This results in 1250, 250, and 250 images for training, validation, and testing, respectively, while keeping the same balance between the day, night, clear, fog, rain and snow images as the original dataset.

Implementation details. All experiments are performed on a A100 GPU. The models are trained with an initial learning rate of 6×10^{-5}. The optimizer is AdamW [17] with epsilon $1e^{-8}$ and weight-decay 0.01 over 200 epochs. For all experiments, the learning rates are scheduled with a polynomial strategy with power 0.9 including 10 warm-up epochs.

The data augmentation includes random horizontal flips, random scaled crops, gaussian blur and random color jitter. The proposed architecture ReCoLaF is built with Segformer [22] encoder-decoders with an MiT-B2 backbone [22]. All Dempster-Shafer-based modules are fully differentiable and integrated end-to-end with standard backpropagation. It is nevertheless noticeable that only the class-wise reliability estimation module contains trainable parameters.

5.2 Experimental Setup

We evaluate the robustness, efficiency, and interpretability of ReCoLaF under challenging multimodal perception conditions. We compare against recent state-of-the-art fusion methods, analyse the estimated reliability scores, and report model efficiency in terms of FLOPs, parameters, and inference time.For both DeLiVER and MUSES datasets, we adopt the protocol introduced by [14], supported by [15], and followed by [5], where all sensor modalities are used during training, and sensor failure scenarios are simulated at inference time.

To simulate sensor failures, we selectively disable one or more modalities during inference by replacing the corresponding input with zero-filled tensors. This strategy, shown to be simple, reproducible, and effective for evaluating fusion robustness in prior work [15], allows us to assess the model's robustness without requiring modality-specific dropout training. For each degraded configuration, we report the mean Intersection over Union (mIoU) [7, 16] over all semantic classes. A modality is considered "available" if its input is provided during inference. Otherwise, it is zero-filled. Results across all tested configurations are presented in Table 1.

5.3 Robustness Analysis

Results on DeLiVER. ReCoLaF outperforms CMNeXt and ECoLaF across most degraded configurations and achieves competitive performance compared to CAFuser. In full-modality settings (all four sensors), CAFuser achieves the highest mIoU (53.87%), followed closely by ReCoLaF (49.90%) and ECoLaF (49.25%). However, ReCoLaF shows stronger robustness in challenging configurations. For instance, in the *RGB + Event* configuration, ReCoLaF scores 35.69%, surpassing ECoLaF (31.44%), CAFuser (26.29%), and CMNeXt (20.37%). Overall, these results highlight the strong dependency of both CMNeXt and CAFuser to the Depth modality. This can be explained by the fact that this modality is never degraded during the training phase and is not realistically impacted by adversarial weather conditions due to the synthetic nature of the dataset. Therefore, the Depth modality is always very informative whereas the RGB modality can be impacted by over-exposure, under-exposure or motion blur during training, encouraging the models to strongly rely on the Depth modality. On average across all configurations, ReCoLaF achieves 34.82% mIoU, outperforming ECoLaF (32.35%), CMNeXt (30.94%), and CAFuser (33.35%).

Results on MUSES. ReCoLaF continues to demonstrate strong robustness against sensor failure under real-world adverse weather conditions. While CAFuser

obtains once more the highest performance in the full-modality setting (72.88%), it shows a strong dependency on the RGB modality, whereas CMNeXt interestingly shows a dependency on the *RGB+LiDAR* and the *RGB+Radar* combinations. Under partial modality configurations, ReCoLaF consistently surpasses other methods. For instance, in the *Event+LiDAR* setup, ReCoLaF achieves 33.90%, compared to 19.49% (ECoLaF), 16.83% (CAFuser), and 2.62% (CMNeXt). In the *RGB+Radar* configuration, ReCoLaF reaches 66.56%, slightly outperforming ECoLaF (65.02%), CAFuser (65.10%), and CMNeXt (61.66%). On average across all configurations, ReCoLaF scores 45.85% mIoU, compared to 41.04% (ECoLaF), 36.38% (CMNeXt), and 42.02% (CAFuser). Regarding the LiDAR modality, there is a clear difference between DeLIVER and MUSES in terms of performances. This may be explained by the fact that the real-world LiDAR images from the MUSES dataset are more dense than the synthetic ones from the DeLIVER dataset, making the classical transformer-based encoder-decoders more effective to extract information from these real-world images.

The presented results confirm that ReCoLaF provides a strong balance between accuracy in favorable conditions and robustness in degraded ones. While CAFuser performs well when all modalities are available, its performance drops more steeply under sensor failures, even though modality-drop was used during training. In contrast, ReCoLaF explicitly limits over-reliance on any single modality by modeling class-specific sensor reliability and incorporating conflict-guided fusion. This design may result in slightly lower peak performance in full-modality settings, but it leads to improved robustness and stability under adverse sensing conditions. Such a trade-off is an intentional design choice, aligned with the requirements of safety-critical autonomous driving systems, where resilience to sensor degradation and failures is often more critical than maximizing accuracy under ideal conditions.

5.4 Sensor Reliability Analysis

To better understand how our model estimates class-specific reliability across semantic classes, we analyze the learned per-modality, per-class reliability scores. Figure 2 presents a radar chart visualization of the estimated reliability scores for four modalities: RGB, Depth, LiDAR, and Event. We observe that Depth consistently shows the highest reliability for structural and planar classes such as *Building*, *Wall*, *Fence*, and *Vegetation*, reflecting its strong geometric representation. In contrast, RGB performs well on texture-rich and visually distinctive classes, such as *Road lines*, *Cars*, and *Traffic signs*. Event cameras show moderate reliability across all classes but do not dominate in any particular category. They tend to follow RGB in shape but with slightly lower scores, indicating their utility in spread but less impactful alone. LiDAR reflects lower reliability in several classes, especially *Traffic signs*, *Wall*, and *Sky*. This may be due to its sparse nature or limitations in vertical resolution for capturing elevated or fine-detailed features. These findings confirm that our class-specific reliability estimation module captures meaningful sensor-class relationships, effectively enabling the fusion framework to weigh sensor contributions adaptively based on semantic content.

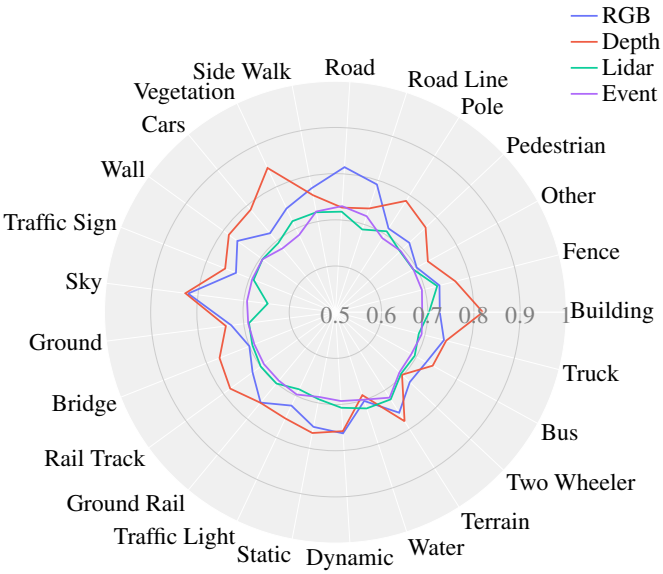

Figure 2: Estimated class-specific reliability across the four sensor modalities of the DeLIVER dataset: RGB, Depth, LiDAR, and Event.

5.5 Ablation study

To evaluate the effectiveness of our two-stage discounting mechanism, we conduct an ablation study comparing ReCoLaF with and without discounting on DeLiVER. Specifically, we remove both the semantic reliability estimation and the conflict-based adjustment, resulting in a version where all modalities contribute equally during fusion. To this end, we train the two models separately.
We compare the following two variants:

- **ReCoLaF w/o discounting:** all mass functions are fused without any reliability-based weighting or conflict-guided discounting. This is equivalent to applying uniform fusion in the Dempster-Shafer framework (i.e., $\beta_k^m = 1$ and $\alpha_m = 1$ in Equations (9) and (11), respectively).
- **ReCoLaF (full):** includes both class-specific reliability estimation and conflict-guided discounting.

As shown in Table 2, removing both discounting stages results in a performance drop of over 6 points in average mIoU. The ablation of the discounting mechanism makes the model highly dependent on the Depth modality, leading to the same lack of robustness as the probabilistic models CMNeXt and CAFuser. This highlights the importance of jointly modeling sensor reliability and inter-modality inconsistency during fusion, particularly under sensor failure conditions.

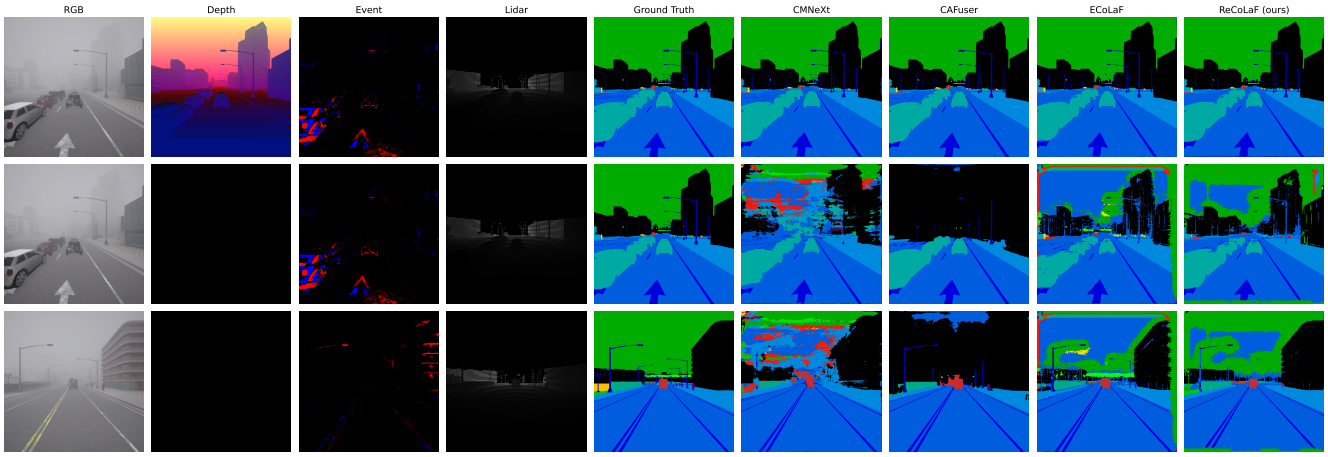

Figure 3: Qualitative segmentation results under degraded conditions (fog and missing depth). Row 1 shows the full-modality scene (RGB, Depth, Event, LiDAR), Row 2 is the same scene under simulated Depth failure (zero-filled), and Row 3 shows a different scene also with missing Depth. From left to right: sensor inputs, ground truth, and predictions from CMNeXt, CAFuser, ECoLaF, and ReCoLaF. ReCoLaF consistently provides cleaner and more accurate segmentations, especially in the presence of sensor failure.

RGB	*Depth*	*Event*	*LiDAR*	ReCoLaF w/o discounting	Full ReCoLaF
✓				12.67	**35.62**
	✓			41.70	**42.06**
		✓		1.99	**3.77**
			✓	1.99	**3.13**
✓	✓			49.81	**49.89**
✓		✓		12.67	**35.69**
✓			✓	12.67	**35.55**
	✓	✓		41.70	**42.48**
	✓		✓	41.69	**42.05**
		✓	✓	1.99	**4.14**
✓	✓	✓		49.81	**49.90**
✓	✓		✓	49.81	**49.90**
✓		✓	✓	12.67	**35.66**
	✓	✓	✓	41.69	**42.58**
✓	✓	✓	✓	49.81	**49.90**
mean				28.18	**34.82**

Table 2: Ablation study on DeLiVER: effect of discounting.

5.6 Qualitative Analysis

To further assess the robustness of the proposed framework under degraded sensing conditions, Figure 3 presents qualitative segmentation results from the DeLiVER dataset under fog and simulated Depth modality failure. The first row shows a scene where all four modalities available. In this full-modality configuration, all methods produce reasonably accurate segmentation maps that are sufficiently clean for perception in autonomous systems. The second and third rows simulate sensor failure by removing the Depth input, which is the modality the models are most dependent on. In both cases, CMNeXt produces highly degraded predictions with strong visual artifacts. CAFuser shows some resilience but provides substantial misclassifications across large areas of the image, particularly in the upper regions of the image, where it often confuses the sky with buildings. ECoLaF partially mitigates the issue thanks to its conflict-guided fusion, but still shows significant artifacts and inconsistencies. ReCoLaF, in contrast, maintains structurally consistent predictions and accurate labeling even in the absence of Depth. This illustrates its ability to effectively adjust the contribution of the remaining modalities based on their reliability and mutual agreement. The results confirm the advantage of our two-stage fusion strategy, which combines learned semantic reliability scores with conflict-guided fusion which provide strong resilience to sensor-level uncertainty and failure.

Overall, these qualitative results support our claim that ReCoLaF is more resilient to sensor failures than attention-based (CMNeXt) or condition-aware (CAFuser) fusion approaches, making it a promising choice for deployment in real-world autonomous driving systems.

5.7 Real-Time Inference and Model Efficiency

Table 3 reports FLOPs, parameters count, and inference time for all evaluated models on DeLiVER. ReCoLaF shares the same architectural complexity as ECoLaF, with nearly equivalent computational cost and parameter count, and is significantly more frugal than CAFuser (which requires condition prediction). CMNeXt is the most lightweight, but achieves lower average mIoU under sensor failures. These results demonstrate that ReCoLaF offers a favorable trade-off between robustness and computational cost, making it suitable for real-time autonomous driving applications. Moreover, the number of parameters in the ReCoLaF architecture can be greatly reduced by adopting a shared backbone strategy as in CAFuser.

	CMNeXt	CAFuser	ECoLaF	ReCoLaF
GFLOPs	65.42	699.12	157.12	157.63
# Params (M)	58.73	75.01	103.16	103.16
Inference time (s/img)	0.17	0.38	0.23	0.24

Table 3: FLOPs, parameters and inference time comparison on the DeLIVER dataset.

6 Conclusion

In this paper, we presented ReCoLaF, a novel fusion framework for robust multimodal semantic segmentation in autonomous driving. ReCoLaF combines class-specific sensor reliability estimation with conflict-guided adjustment to adaptively fuse heterogenous sensor modalities under uncertain or degraded conditions. Grounded in Dempster-Shafer theory, our two-stage fusion strategy first discounts each modality based on learned per-class semantic reliability, then further adjusts its contribution based on disagreement with other modalities at the pixel level.

Experiments on the DeLiVER and MUSES datasets demonstrate that ReCoLaF consistently improves robustness compared to strong baselines, including middle fusion approaches with cross-attention mechanism and conflict-only evidential methods, particularly under sensor degradation and failure scenarios. While some competing methods achieve higher performance when all modalities are available, ReCoLaF is intentionally designed to favor robustness and stability under adverse conditions, reflecting a trade-off that is well aligned with the requirements of safety-critical autnomous driving systems. These results highlight the importance of explicitly modeling both the reliability and agreement of sensor modalities to achieve robust scene understanding in autonomous driving systems.

In the current formulation, class-wise sensor reliability scores are learned as global parameters and remain fixed during inference. This design choice provides stable semantic priors while keeping the fusion process tractable and interpretable, and is complemented by a spatially adaptive conflict-based mechanism that locally adjusts modality contributions. Future work will investigate dynamic and pixel-level reliability estimation to better capture region-specific and context-dependent sensor degradation. In addition, parameter-efficient architectures, such as shared backbone designs, will be explored to improve scalability to larger numbers of modalities and higher-resolution inputs, as well as to enhance real-time performance for practical autonomous driving applications.

acknowledgement

The authors are supported by the French National Research Agency (ANR) under grants HAISCoDe, INARI, and EviDeep. This project was provided with computing AI and storage resources by GENCI at IDRIS thanks to the grant 2024-AD011014391 on the supercomputer Jean Zay's A100 partition along with computing resources of CRIANN (Normandy, France).

References

[1] Alireza Asvadi, Luis Garrote, Cristiano Premebida, Paulo Peixoto, and Urbano J Nunes. Multimodal vehicle detection: fusing 3d-lidar and color camera data. *Pattern Recognition Letters*, 115:20–29, 2018.

[2] Tim Brödermann, David Bruggemann, Christos Sakaridis, Kevin Ta, Odysseas Liagouris, Jason Corkill, and Luc Van Gool. Muses: The multi-sensor semantic perception dataset for driving under uncertainty. In *European Conference on Computer Vision*, pages 21–38. Springer, 2024.

[3] Tim Brödermann, Christos Sakaridis, Yuqian Fu, and Luc Van Gool. Cafuser: Condition-aware multimodal fusion for robust semantic perception of driving scenes. *IEEE Robotics and Automation Letters*, 10(4):3134–3141, 2025.

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

[5] Lucas Deregnaucourt, Hind Laghmara, Alexis Lechervy, and Samia Ainouz. A conflict-guided evidential multimodal fusion for semantic segmentation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1373–1382, February 2025.

[6] Jean Dezert and Florentin Smarandache. A new probabilistic transformation of belief mass assignment. *CoRR*, abs/0807.3669, 2008.

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[8] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020.

[9] Jannik Fritsch, Tobias Kühnl, and Andreas Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 1693–1700, 2013.

[10] Mihreteab Negash Geletu, Dănuţ-Vasile Giurgi, Thomas Josso-Laurain, Maxime Devanne, Mengesha Mamo Wogari, and Jean-Philippe Lauffenburger. Evidential deep learning-based multi-modal environment perception for intelligent vehicles. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–6. IEEE, 2023.

[11] Ling Huang, Su Ruan, Pierre Decazes, and Thierry Denœux. Deep evidential fusion with uncertainty quantification and reliability learning for multimodal medical image segmentation. *Information Fusion*, 113:102648, 2025.

[12] Anne-Laure Jousselme, Dominic Grenier, and Éloi Bossé. A new distance between two bodies of evidence. *Information fusion*, 2(2):91–101, 2001.

[13] Bingyu Li, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation, 2024.

[14] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19800–19808, 2022.

[15] Chenfei Liao, Kaiyu Lei, Xu Zheng, Junha Moon, Zhixiong Wang, Yixuan Wang, Danda Pani Paudel, Luc Van Gool, and Xuming Hu. Benchmarking multimodal semantic segmentation under sensor failures: Missing and noisy modality robustness, 2025.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[18] Arnaud Martin, Anne-Laure Jousselme, and Christophe Osswald. Conflict measure for the discounting operation on belief functions. In *2008 11th International Conference on Information Fusion*, pages 1–8, 2008.

[19] Glenn Shafer. *A mathematical theory of evidence*. Princeton University Press, 1976.

[20] Zheng Tong, Philippe Xu, and Thierry Denœux. Fusion of evidential cnn classifiers for image classification. In *International Conference on Belief Functions*, pages 168–176. Springer, 2021.

[21] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 128(5):1239–1285, 2020.

[22] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[23] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *CVPR*, 2023.

[24] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023.