A Conflict-Guided Evidential Multimodal Fusion for Semantic Segmentation

Lucas Deregnaucourt[1] Hind Laghmara[1] Alexis Lechervy[2] Samia Ainouz[1]

[1]INSA Rouen Normandie, Univ Rouen Normandie, Université Le Havre Normandie,
Normandie Univ, LITIS UR 4108, F-76000 Rouen, France
{lucas.deregnaucourt, hind.laghmara, samia.ainouz}@insa-rouen.fr

[2]GREYC, Normandie Univ, Unicaen, Ensicaen, UMR CNRS 6072
alexis.lechervy@unicaen.fr

Abstract

This article presents a novel and robust approach to semantic segmentation based on the fusion of different image modalities (conventional and non-conventional images). The robustness of fusion methods and their ability to tolerate sensor failures are crucial challenges for their deployment in real-world environments. It is essential to develop unique fusion models that can operate even in the absence of certain modalities during inference. However, current fusion methods have a strong dependence on the RGB branch, resulting in significant performance losses in case of its unavailability. To address this issue we propose ECoLaF (Evidential Conflict-guided Late Fusion), a 'late fusion' method based on Dempster-Shafer theory. This method adaptively reduces the output of each modality according to their conflicts before fusing them. Experimental results show that our approach outperforms state-of-the-art methods in terms of robustness on the MCubeS and DeLiVER datasets, especially when the RGB sensor is not operational. This study offers new perspectives for improving the robustness of semantic segmentation in multimodal contexts. Code is available at https://github.com/deregnaucourtlucas/ECoLaF.

1. Introduction

Road scene analysis is a fundamental task for autonomous driving systems. To move safely, an autonomous system should thoroughly analyze and understand the environment in which it navigates. Advances in computer vision continuously attempt to propose the best solutions to improve the vehicle's perception. Besides road scene object detection task which gains a huge interest recent years [14], semantic segmentation [18] is still one of the most interesting ways for analyzing and understanding the road scene.

The advent of fully connected networks significantly boosted interest in semantic segmentation [1, 3, 15, 23]. These networks deliver outstanding performances due to the ability of convolutional filters to effectively capture local information. Recently, the advent of transformers in semantic segmentation architectures [35] pushed considerably the performances, as local patches more effectively capture global information in the input image compared to convolutional filters. However, the ambiguity between the classes is still an open problem and can lead to misclassifications. Indeed, the probabilistic theorical framework typically used to train neural networks tends to make them overconfident and by so unable to express any kind of uncertainty [12]. In autonomous driving applications, an overconfident model that can strongly fail in some situations is hardly trustworthy and therefore might never be used for real-life applications.

One promising approach to overcome the ambiguity issue is provided by the Dempster-Shafer theory (DST) that allows the models to express themselves not only on single classes but also on set of classes [4, 29]. The power of this theory is its ability to model and reason about imprecise and uncertain problems, and has more obvious advantages in the representation and combination of uncertain or incomplete information. Moreover, in recent years, neural networks based on DST were proposed to better model the uncertainty [6,28,37], showing an improved handling of the classes ambiguity.

Another key challenge related to semantic segmentation in autonomous driving applications is the robustness of the model while facing uncommon scenarios such as bad weather conditions [44] or sensor failure [10]. This is the reason why attention was drawn to multimodal fusion to overcome the weaknesses of the RGB modality as a standalone source. Nevertheless, most of the works only consider fusing two modalities [11,27,33], which is not enough to cover every real-life situation. Recent works focused on adding even more modalities [2,20,43] to push forward the complementarity of sensors and thus the performances in perception. Despite its outstanding performances on the challenging MCubeS [20] and DeLiVER [43] datasets, the state-of-the-art architecture CMNeXt [43] based on the *middle fusion* strategy drops from 51.54% in mIoU to 5.05% when the RGB images are not available on the MCubeS dataset and from 53.01% to 20.54% when the depth images are not available on the DeLiVER dataset. This lack of robustness clearly shows that the *middle fusion* strategy almost exclusively relies on one modality. Otherwise, another fact that could drop the performances in using multimodality comes from a discordance between some modalities prior to fusion. This can be highly problematic in real-life applications such as autonomous driving where human lives are at stake.

To address this robustness issue, to overcome the *middle fusion* strategy issue and the conflict between modalities, we propose a *late fusion* method based on Dempster-Shafer theory using the distance-based conflict between the mass functions [26]. This method adaptively weakens the output of each modality according to their conflicts before fusing them. This fusion module is purely conflict-oriented and therefore requires no additional parameters to be trained, making it adaptive and completely agnostic to the choosen encoder-decoder architecture. The idea is that if a modality is highly in conflict with the others given a certain pixel, it is reasonable to consider that either the model associated with this modality struggles to classify this pixel or there is a sensor failure. Therefore, it is suitable to weaken the mass function of this conflicting modality at this particular pixel according to its conflict in order to make the information fusion easier.
By averaging the mIoU on all combinations of sensors, our method surpasses the state-of-the-art model by $+13.01\%$ on MCubeS and $+1.41\%$ on DeLiVER. Despite not showing the best performances when every sensor is operational our method proves to be a good trade off between performances and robustness.

2. Related Work

2.1. Semantic segmentation

Since the introduction of fully convolutional networks (FCN) [1,3,15,23], computer vision tasks, in particular semantic segmentation, have seen significant progress. Recently, the emergence of attention mechanisms in computer vision architectures [8] pushed forward even further the performances. Indeed, Vision Transformer models [1,22] tackled the main issue of the FCN which is their poor capability to extract global features. Despite these improvements, the models using only RGB images struggle to segment correctly complex scenes where the contrast is low or when the weather conditions are not optimal (night, fog, rain, snow).

2.2. Multimodal semantic segmentation

To overcome the weaknesses of RGB images, attention has turned to the addition of complementary modalities [45] such as thermal [30, 38], depth [19, 27] and polarization [17, 39]. Most of the existing methods explore the *middle fusion* strategy. In MCubeSNet [20], the feature maps of each separated backbone are concatenated and passed through a convolutional region-guided filter selection layer. CMX [42] and CMNeXt [43] propose a highly scalable fusion framework with multi-level cross-modal interactions. The main issue with most of these works is their insufficient robustness and adaptability, primarly because they essentially rely on the RGB branch and maintain fixed weights for fusion layers at the end of the training.

2.3. Evidential neural networks

The distance-based evidential neural network classifier proposed in [5] paved the way to incorporate the Dempster-Shafer Theory(DST) in neural network architectures [13, 28, 41]. This work was extended to deep neural networks for classification and semantic segmentation tasks [36, 37]. However, these attempts have been confined to relatively small and well-structured datasets. The primary impediment has been the algorithmic complexity of DST, which scales exponentially with the size of the frame of discernment Ω, containing 2^K subsets where $K = |\Omega|$ is the number of classes in our case. To this end, [6] proposes algorithmic optimizations to render the evidential networks highly scalable.

3. Dempster-Shafer Theory

3.1. Background on belief functions

Dempster-Shafer theory [4, 29] is a generalized mathematical framework making it possible to reason about uncertain problems. This framework also has advantages in the representation and fusion of uncertain or incomplete information.

Let consider the *frame of discernment* Ω as a finite set of variables ω which refers to K elementary events to a given problem ($\Omega = \{\omega_1, \omega_2, ..., \omega_K\}$). For classification or semantic segmentation tasks, the elementary events are the dataset classes. The power set of Ω is defined as the set of all the 2^K possible subsets of Ω. It is presented as follows:

$$2^\Omega = \{\emptyset, \{\omega_1\}, ..., \{\omega_k\}, \{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}, ..., \Omega\} \quad (1)$$

where the $\{w_i\}$ elements are called singletons and $\emptyset$ denotes the empty set.

The main point of the Dempster-Shafer theory is the possibility of representing partial knowledge of the value of ω with the basic belief assignment (*bba*). A *bba* is a function m from 2^Ω to $[0,1]$ defined as follows:

$$\begin{aligned} m : 2^\Omega &\to [0,1] \\ A &\mapsto m(A) \end{aligned} \tag{2}$$

where m satisfies the following constraint:

$$\sum_{A \in 2^\Omega} m(A) = 1 \tag{3}$$

An element A of Ω is called a *focal element* when $m(A)>0$, and the set containing all these elements is called a *body of evidence* (BOE) or *focal set*.

3.2. Conflict measure and mass discounting

In Dempster-Shafer theory, the conflict can be represented by the contradiction between mass functions. Therefore, two experts (*e.g.* modalities, sensors,...) are in conflict if they are far from each other in the *bba*'s space. To this end, [25] propose a distance-based definition of the conflict between two mass functions m_1 and m_2 (Eq. (4)). F_1 and F_2 are respectively the focal sets of m_1 and m_2 and $|.|$ denotes the cardinality function.

$$Conf_{1,2} = \left(1 - \frac{1}{|F_1||F_2|} \times \sum_{X \in F_1} \sum_{Y \in F_2} 1_{\{X \subseteq Y\}}\right) \times d_{1,2} \tag{4}$$

where $d_{1,2}$ is the Jousselme distance [16] between m_1 and m_2:

$$d_{1,2} = \sqrt{\frac{1}{2}(m_1 - m_2)^T \underline{D}(m_1 - m_2)} \tag{5}$$

where $\underline{D}$ is a $2^{|\Omega|} \times 2^{|\Omega|}$ matrix defined as follows:

$$\underline{D}(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad \forall A, B \in 2^\Omega \tag{6}$$

The conflict associated with the mass function m_i is defined by:

$$Conf_i = \frac{1}{M-1} \sum_{j=1, i \neq j}^{M} Conf_{i,j} \tag{7}$$

where M is the number of experts. This conflict allows to assess the reliability α of each expert which can be used to weaken the *bbas* by the discounting procedure before fusing them:

$$\begin{cases} m^\alpha(X) = \alpha m(X) \\ m^\alpha(\Omega) = 1 - \alpha(1 - m(\Omega)) \end{cases} \tag{8}$$

A way to compute the reliability α_i from the conflict $Conf_i$ is proposed in [26]:

$$\alpha_i = (1 - Conf_i^\lambda)^{\frac{1}{\lambda}} \tag{9}$$

with $\lambda > 0$.

3.3. Information fusion

As mentioned in Sec. 3.1, the fusion of information is a central feature of the DST. The most common way to combine two *bbas* m_1 and m_2 defined on the same frame of discernment Ω is the Dempster's rule [29], denoted here by $\oplus$. It is defined by $m_{DS}(\emptyset) = 0$ and $\forall A \in 2^\Omega \backslash \{\emptyset\}$ by:

$$m_{DS}(A) = (m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{\substack{B \cap C = A \\ B, C \in 2^\Omega}} m_1(B) m_2(C) \tag{10}$$

where κ represents the degree of conflict between the two *bbas* defined by:

$$\kappa = \sum_{\substack{B \cap C = \emptyset \\ B, C \in 2^\Omega}} m_1(B) m_2(C) \tag{11}$$

This fusion can be seen as the normalized version of the conjunctive rule [31] which is defined by:

$$m_\cap(A) = \sum_{\substack{B \cap C = A \\ B, C \in 2^\Omega}} m_1(B) m_2(C) \tag{12}$$

3.4. Probability transformation

For semantic segmentation task, the decision is made among elements of the *frame of discernment*. However a non-zero mass can be assigned to a set of disjunctive classes, making it harder to make a precise decision. Therefore, it is required to transform the *bbas* into probabilities by redistributing the partial conflicts among the singletons. The most common transformation, proposed in [31], is the pignistic probability transformation denoted by $BetP(.)$:

$$BetP(\omega_k) = \sum_{\omega_k \in A \subseteq \Omega} \frac{m(A)}{|A|}, \quad \forall \omega_k \in \Omega \tag{13}$$

This transformation redistributes uniformly the partial conflict $m(A)$ among the singletons $\omega_k \in A \subseteq \Omega$. A generalized pignistic transformation was proposed in [7]:

$$DSmP_\varepsilon(\omega_k) = \sum_{A \in 2^\Omega} m(A) \frac{\sum_{a \in \omega_k \cap A} m(a) + \varepsilon \cdot |\omega_k \cap A|}{\sum_{a \in A} m(a) + \varepsilon \cdot |A|} \tag{14}$$

where $\varepsilon > 0$ is a parameter that controls the effect of element's cardinality in the partial conflict redistribution.

4. Proposed framework

To achieve evidential multimodal semantic segmentation, the proposed ECoLaF architecture is based on a classical *late fusion* approach along with Dempster-Shafer theorical framework presented in Sec. 3 where each encoder-decoder outputs mass functions instead of probabilities. We

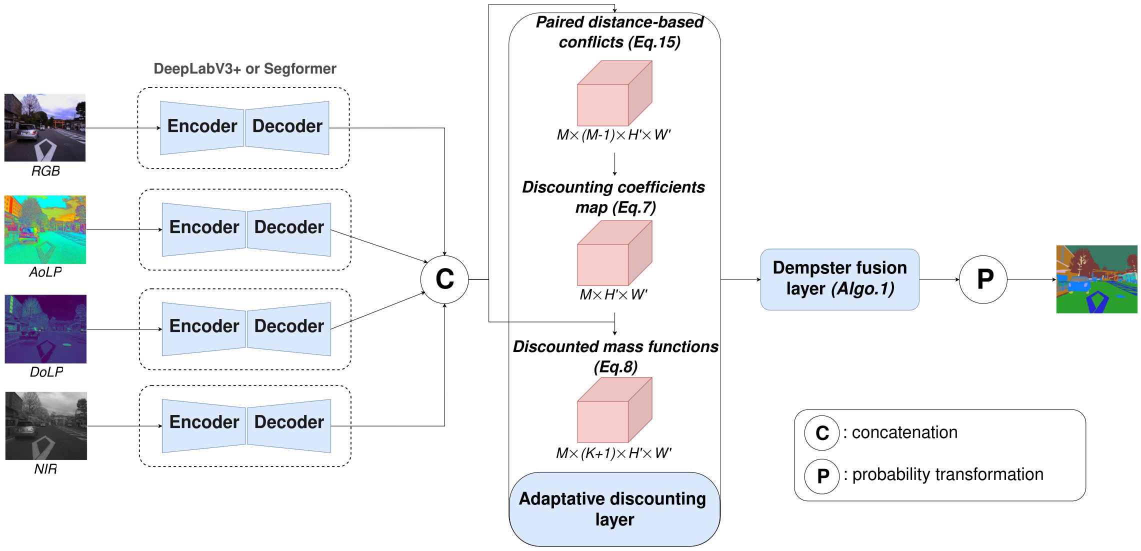
DeepLabV3+ or Segformer
Encoder
Decoder
RGB
AoLP
DoLP
NIR
C
Paired distance-based conflicts (Eq.15)
M×(M-1)×H'×W'
Discounting coefficients map (Eq.7)
M×H'×W'
Discounted mass functions (Eq.8)
M×(K+1)×H'×W'
Adaptative discounting layer
Dempster fusion layer (Algo.1)
P
C : concatenation
P : probability transformation

Figure 1. **ECoLaF architecture**. Each modality is associated with an independant encoder-decoder model such as DeepLabV3+ [3] or Segformer [40]. The mass functions of each modality are weakened by the adaptative discounting layer and fused by the Dempster's rule (Eq.10). The final decision is made after converting the mass functions into probabilities (Eq.14).

describe the evidential encoder-decoders in Sec 4.1, the adaptative discounting layer in Sec. 4.2 and the fusion and decision making process in Sec. 4.3.

4.1. Evidential encoder-decoders

In Figure 1, each of the M modalities is associated with an independent encoder-decoder network. Following the recommendations of [5] to apply Dempster-Shafer theory to neural networks, we reduce the possible focal elements of the mass functions to singletons $\omega_k \in \Omega$ and the frame of discernment Ω itself where ω_k corresponds to the kth class of the dataset and Ω corresponds to the set of all classes with $K = |\Omega|$. Each evidential encoder-decoder is given an image of shape $C_m \times H \times W$ where C_m corresponds to the number of channels of the $m\text{-}th$ modality. Each encoder-decoder outputs a *mass functions maps* of shape $(K+1) \times H' \times W'$ by applying a *softmax* to the last *feature maps* of the decoder. A probabilistic encoder-decoder can thus be seen as a particular evidential one that always outputs $m_{ij}(\Omega) = 0 \quad \forall (i,j) \in [\![1, H']\!] \times [\![1, W']\!]$.
The *mass functions maps* are then concatenated and weakened by the adaptative discounting layer presented in the next section.

4.2. Adaptative discounting layer

We present here a method to discount the mass functions of each modality pixel-wise depending on their conflict with the others. This method is purely conflict-guided and requires no additional parameters to be trained, making it highly adaptative. Given that we only consider singletons and Ω to construct the mass functions, the cardinality of their focal sets is equal to $K+1$: $m(\omega_1), \dots, m(\omega_K), m(\Omega)$. Equation (4) can be simplified as follows:

$$Conf_{1,2} = \left(1 - \frac{2K+1}{(K+1)^2}\right) \times d_{1,2} \tag{15}$$

Moreover, the $\underline{D}$ matrix defined in the Eq. (6) is now reduced to a $(K+1) \times (K+1)$ matrix:

$$\underline{D}(A,B) = \begin{cases} 1 & \text{if } A = B \\ \frac{1}{K} & \text{if } A = \Omega \text{ xor } B = \Omega \\ 0 & \text{else} \end{cases} \tag{16}$$

which leads to a computational simplification of the Eq. (5):

$$d_{1,2} = \sqrt{\frac{1}{2} S_{1,2}} \tag{17}$$

where
$$S_{1,2} = \sum_{A \in \Omega \cup \{\Omega\}} (m_1(A) - m_2(A))^2 + \frac{2}{K}(m_1(\Omega) - m_2(\Omega)) \times \sum_{A \in \Omega} (m_1(A) - m_2(A)).$$

Given an element $(i,j) \in [\![1, H']\!] \times [\![1, W']\!]$, the conflicts between the M *mass functions' maps* are computed pair-wise following Eq. (15). The conflicts associated with each modality are then computed following the Eq. (7). The discounting factors are finally obtained with the Eq. (9). We choose here $\lambda = 2$ for the good mathematical properties of the resulting function linked to the l^2 norm, as explained in [26]. By doing so, the more a mass function is in conflict with the others, the more it will be discounted.

Following the Equation (8), the resulting *discounting maps* of shape $M \times H' \times W'$ are used to discount the *mass functions maps* of each modality element-wise. The idea is that if a modality is highly in conflict with the others given a certain pixel, it is reasonable to consider that either the model associated with this modality struggles to classify this pixel or there is a sensor failure. Therefore, it is suitable to weaken the mass function of this conflicting modality at this particular pixel according to its conflict in order to ease the information fusion. The discounted *mass functions maps* are then fused using the Dempster's rule.

4.3. Dempster fusion layer and decision making

Following the Equation (10) the discounted *mass functions maps* of each modality are fused element-wise, leading to a final *mass functions maps* of shape $(K+1) \times H \times W$. Algorithm 1 shows the computationally optimized Dempster's rule proposed in [6]. In our case, M is the number of modalities and K is the number of classes.
The decision is made by transforming the mass functions into probabilities following the Eq. (14). We choose here $\varepsilon = 0.001$ as recommended in [7] since the smaller ε, the bigger probability information content value [32], facilitating the decision making.

Algorithm 1 Scalable Dempster's rule (Eq.10)

```
Require: $M$ mass functions $m_1, \dots, m_M$
  $m_\cap(\Omega) = \prod_{i=1}^{M} m_i(\Omega)$
  for $j = 1, \dots, K$ do
    $m_\cap(\{\omega_j\}) = \prod_{i=1}^{M} (m_i(\{\omega_j\}) + m_i(\Omega)) - m_\cap(\Omega)$
  end for
  return $m_{DS}(.) = \frac{m_\cap(.)}{\sum_{A \in \{\omega_1,\dots,\omega_K,\Omega\}} m_\cap(A)}$
```

5. Experiments

5.1. Datasets and implementation details

MCubeS [20] is an outdoor dataset which contains paired images of four modalities, namely RGB, Angle of Linear Polarization (AoLP), Degree of Linear Polarization (DoLP) and Near-Infrared (NIR), to study semantic material segmentation of 20 classes. It has 302/96/102 image pairs for training/validation/testing at the size of 1224×1024.
DeLiVER [43] is a synthetic dataset which contains paired images of four modalities, namely RGB, Depth, Event and LiDAR in various weather conditions along with sensor failure scenarios, namely motion blur, over-exposure, under-exposure, LiDAR-jitter and Event low-resolution, to study semantic segmentation of 25 classes. It has 3983/2005/1897 front-view image pairs for training/validation/testing at the size of 1042×1042.

Implementation details. All experiments are performed on a A100 GPU. We train our models based on DeepLabV3+ encoder-decoder with an initial learning rate of 0.05. The optimizer is SGD [34] with momentum 0.9 and weight-decay $5e^{-4}$. The models based on Segformer encoder-decoder are trained with an initial learning rate of $6e^{-5}$. The optimizer is AdamW [24] with epsilon $1e^{-8}$ and weight-decay 0.01. The data augmentation includes random horizontal flips and random scaled crops for the MCubeS dataset along with random gaussian blur and random color jitter for the DeLiVER dataset. The models are respectively trained over 500 and 200 epochs with a batch size of 8 and 16 on the MCubeS and the DeLiVER datasets. For all experiments, the learning rates are scheduled with a polynomial strategy with power 0.9 including 10 warm-up epochs. We use cross-entropy loss function.

Experimental setup. The models are trained with every available modalities once, meaning that we don't retrain a new model for every combination of sensors. To evaluate the robustness of the models, we follow the same experimental setup as in [20]: when a certain modality is excluded during testing, the encoder is fed with a zeroed-out image for that modality to simulate a sensor failure. To ensure fair comparison, we load the provided trained MCubeSNet [20] and CMNeXt [43] models with all modalities and apply the aforementioned experimental setup.

5.2. Robustness comparison against state-of-the-art methods

To verify the efficiency of the proposed ECoLaF architecture in terms of robustness, we evaluate it on the challenging MCubeS [20] and DeLiVER [43] datasets with every combination of sensors and compare it against the state-of-the-art. We use the *mean Intersection over Union* (mIoU) metric [9, 21] to evaluate the models.
Results on MCubeS Table 1 summarizes robustness comparison between our ECoLaF method, MCubeSNet [20] and CMNeXt [43] on MCubeS dataset. A striking observation is that both MCubeSNet and CMNeXt performances heavily fall down when the RGB sensor is not available. Therefore, despite showing the best performances when every sensor is available, CMNeXt can't be considered robust and adaptative. Indeed, our ECoLaF-DeepLabV3+ outperforms it by $+13.01\%$ in mIoU when we average the performances on all possible scenarios. Another interesting observation is that our transformer-based ECoLaF-Segformer is still strongly affected by the RGB sensor failure. This is essentially due to the fact that it is very hard to train a transformer-based encoder-decoder from scratch with only 302 images.
The per-class detailed performances of the ECoLaF-DeepLabV3+ is shown in Table 2. We can see that the

				convolution-based models		transformers-based models	
RGB	AoLP	DoLP	NIR	MCubeSNet [20]	ECoLaF-DeepLabV3+(ours)	CMNeXt [43]	ECoLaF-Segformer(ours)
✓				30.79	43.49	42.32	**46.48**
	✓			3.63	**21.45**	2.1	10.45
		✓		1.66	**35.44**	3.42	19.84
			✓	1.00	**32.81**	2.15	16.79
✓	✓			38.10	43.36	**48.81**	46.48
✓		✓		35.98	44.95	**49.00**	48.11
✓			✓	33.16	44.39	**48.36**	45.01
	✓	✓		4.60	**36.35**	1.43	27.61
	✓		✓	1.67	**36.81**	1.74	23.14
		✓	✓	1.12	**41.53**	3.15	27.19
✓	✓	✓		41.54	45.26	**49.06**	48.75
✓	✓		✓	40.61	44.25	**49.78**	47.77
✓		✓	✓	39.53	45.57	**50.02**	49.85
	✓	✓	✓	2.74	**41.72**	5.05	33.31
✓	✓	✓	✓	43.26	45.74	**51.54**	49.85
mean				21.29	**40.21**	27.20	36.04

Table 1. Performances comparison of using different modalities in mIoU(%) on MCubeS dataset. Bold values represent the best performances to the nearest rounding for each combination of modalities.

RGB	AoLP	DoLP	NIR	asphalt	concret	metal	road marking	fabric	glass	plaster	plastic	rubber	sand	gravel	ceramic	cobbles	brick	grass	wood	leaf	water	human body	sky	mean
✓				78.7	41.0	44.3	65.2	**20.5**	44.2	**4.1**	24.3	20.9	64.3	60.9	20.2	63.0	37.1	**55.0**	37.7	71.9	21.2	0	95.3	43.5
	✓			63.8	32.6	32.9	0	0	0	1.7	0	0	5.2	12.5	0	39.6	31.4	24.9	0	61.9	**35.4**	0	87.2	21.4
		✓		73.0	35.6	39.1	44.1	13.0	41.8	1.3	24.1	26.2	53.9	33.5	22.4	50.1	33.9	26.3	31.8	63.8	0	0	94.8	35.4
			✓	74.4	26.9	40.4	61.1	0	40.9	**4.0**	0	0	60.5	47.3	0	57.7	30.0	42.5	35.3	66.8	0	14.3	54.2	32.8
✓	✓			79.1	41.4	45.1	60.3	19.3	41.1	3.6	23.0	14.3	64.9	61.3	18.3	63.6	38.6	53.7	37.0	73.0	34.1	0	95.4	43.4
✓		✓		**80.4**	42.2	47.3	65.7	20.1	49.3	2.7	**28.3**	**29.1**	64.7	59.9	**24.7**	64.9	38.9	**55.1**	40.4	73.1	16.0	0	**96.0**	44.9
✓			✓	79.3	43.1	46.7	66.7	19.0	47.3	**4.1**	23.1	15.5	**66.7**	**62.5**	18.7	65.4	35.6	54.1	41.9	74.2	13.2	16.0	95.4	44.4
	✓	✓		71.2	38.3	41.2	36.3	9.4	38.5	1.4	21.1	25.9	56.5	28.0	21.2	46.8	36.2	30.4	30.6	67.3	31.7	0	95.0	36.3
	✓		✓	73.9	39.5	41.8	56.4	0	37.6	3.2	0	0	60.6	47.4	0	54.7	32.8	42.4	34.4	70.0	29.2	19.9	92.3	36.8
		✓	✓	76.3	41.3	45.3	61.6	9.5	47.5	2.0	21.3	27.6	62.2	46.6	22.1	62.4	34.4	43.8	39.1	71.1	0	**21.0**	95.4	41.5
✓	✓	✓		79.9	43.1	48.0	61.9	19.0	47.9	2.4	26.6	27.6	65.3	60.5	24.3	64.0	**40.1**	54.1	39.3	73.8	31.3	0	**96.0**	45.3
✓	✓		✓	79.2	43.9	47.6	65.3	17.2	46.2	3.7	19.7	9.5	66.4	62.2	16.4	65.9	38.4	53.7	41.2	74.2	30.5	7.9	95.6	44.2
✓		✓	✓	79.7	44.1	49.1	**67.6**	18.5	**50.7**	2.8	26.7	28.3	**66.8**	61.5	24.4	66.4	38.0	54.4	**42.6**	74.3	11.1	8.5	**96.0**	45.6
	✓	✓	✓	75.5	42.2	46.1	58.0	7.5	46.3	1.9	17.4	22.1	62.9	46.9	19.6	58.4	36.8	43.6	38.6	72.6	26.1	16.2	95.5	41.7
✓	✓	✓	✓	79.6	**44.6**	**49.6**	66.2	17.2	50.2	2.6	23.6	25.2	**66.7**	61.4	23.5	**66.5**	39.4	54.0	41.8	**74.7**	27.0	4.8	**96.0**	**45.7**

Table 2. Results of ECoLaF-DeeplabV3+ on MCubeS in per-class IoUs(%). Bold values represent the best performances to the nearest rounding for each combination of modalities.

per-class best performances are not always reached when all the sensors are available, showing the negative impact of the classes ambiguity when fusion multiple experts. The most striking example is the class *human body*. When looking at the single-modality performances, it seems that the NIR modality is the only one to be able to detect human body. However, when looking at the two-modalities performances, we can clearly observe an improvement when the NIR modality is combined with any of the others which shows their complementarity to better handle the ambiguity for the *human body* class. Nevertheless, when the NIR modality is available with at least two others the performances drop. It means that those other modalities agree to confuse the *human body* pixels with other classes, increasing the class ambiguity instead of decreasing it. Figure 2 and Figure 3 respectively show qualitative results when all the modalities are available and when the RGB modality is partially blackened. The difference of MCubeSNet and CMNeXt predictions between Fig. 2 and Fig. 3 clearly highlights their sensitivity to RGB sensor failures. On the other hand, the proposed ECoLaF-DeepLabV3+ is only slightly impacted by this sensor failure. This shows that ECoLaF does not exclusively relies on the RGB modality and is able to extract and fuse enough information from the non-conventional modalities to output a satisfying semantic segmentation mask.

Results on DeLiVER Table 3 summarizes robustness comparison between our ECoLaF method and CMNeXt [43] on DeLiVER dataset. The experiments are not carried out with MCubeSNet [20] since this method needs semantic segmentation masks that are not provided in the DeLiVER dataset. This time, the transformers-based ECoLaF-Segformer outperforms both CMNeXt and the convolution-based ECoLaF-DeepLabV3+ in terms of robustness. This can be explained by the diversity and the amount of training images, namely 3983, compared to MCubeS making it easier for transformers-based architectures to converge. When looking at the single-modality performances, we can see that our models better distribute the information among the RGB and the Depth modalities whereas CMNeXt almost exclusively relies on the Depth images. Therefore, even our ECoLaF-DeepLabV3+ surpasses the transformers-based CMNeXt when the Depth sensor is out.

In the light of these experiments, it seems that the Event and LiDAR modalities are too weakly informative to make a significant contribution to improving performances. Moreover, these modalities can't insure on their own a minimum level of performances. These kind of observations are very important to take into account while building perception systems since some sensors can work alone and some cannot but may be useful by bringing complementary information.

5.3. Ablation studies

To attest the contribution of the adaptive discounting layer to robustness, we carry out performances comparison of ECoLaF-DeepLabV3+ and ECoLaF-Segformer with and without this layer. The performances on the MCubeS and DeLiVER datasets are respectively summarized in Tab. 4 and Tab. 5.

By averaging the performances over all the possible combinations of sensors, the addition of the adaptive conflict layer respectively increased the performances of ECoLaF-DeepLabV3+ and ECoLaF-Segformer by $+9.96\%$ and $+8.90\%$ on MCubeS, and $+5.55\%$ and $+4.17\%$ on DeLiVER in mIoU. It is striking that the models trained without the discounting layer behave similarly to CMNeXt and MCubeSNet by almost exclusively taking the information from one modality, namely RGB on the MCubeS dataset and Depth on the DeLiVER dataset. Therefore, their performances heavily drop when the RGB and Depth sensors are respectively out on MCubeS and DeLiVER datasets.

6. Conclusion

In this work, we tackle the issue of robustness in semantic segmentation task for road scene analysis application. We propose a *late fusion* approach based on Dempster-Shafer theory which adaptively discounts the information provided by each modality depending on its conflict with the others. Our ECoLaF method shows a great resistance to sensor failures even when RGB images are not available, achieving overall best performances on the MCubeS and DeLiVER datasets considering all possible combinations of available sensors. On the other hand, the state-of-the-art CMNeXt strongly struggles to maintain satisfying performances when the most informative sensor fails, namely RGB on the MCubeS dataset and Depth on the DeLiVER dataset. ECoLaF better distributes the information among the modalities, making it more robust to sensor failures. In order to remediate the performances drop of all-modality fusion in some cases compared to the state-of-the-art (*e.g.* when the RGB modality is available), we will focus on the inclusion of additional sub-sets as focal elements in order to better model the ambiguity in decision-making.

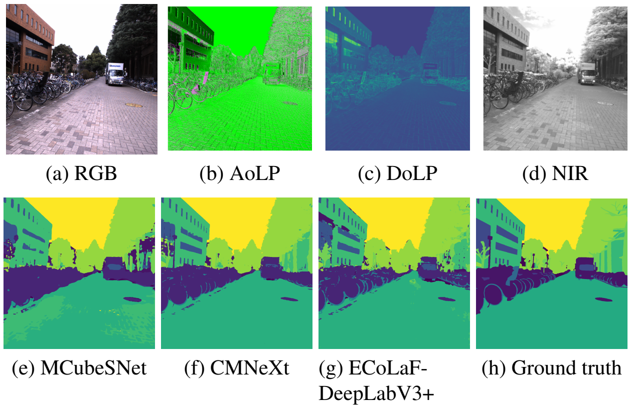
(a) RGB (b) AoLP (c) DoLP (d) NIR
(e) MCubeSNet (f) CMNeXt (g) ECoLaF-DeepLabV3+ (h) Ground truth

Figure 2. Predictions on MCubeS with all modalities available.

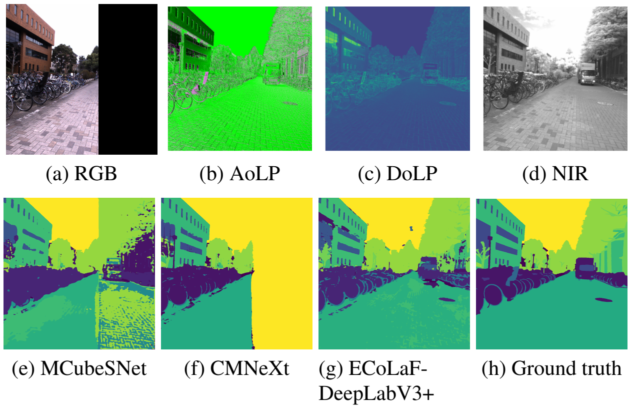
(a) RGB (b) AoLP (c) DoLP (d) NIR
(e) MCubeSNet (f) CMNeXt (g) ECoLaF-DeepLabV3+ (h) Ground truth

Figure 3. Predictions on MCubeS with partial RGB failure.

7. Acknowledgement

The first author is supported by the French National Research Agency (ANR) and Region Normandie under grant HAISCoDe. This project was provided with computing AI and storage resources by GENCI at IDRIS thanks to the

RGB	Depth	Event	LiDAR	CMNeXt [43]	ECoLaF-DeepLabV3+(ours)	ECoLaF-Segformer(ours)
✓				20.62	25.20	**31.44**
	✓			40.29	**40.35**	38.77
		✓		**2.82**	0.39	1.87
			✓	**2.76**	0.12	2.40
✓	✓			**52.96**	46.25	49.23
✓		✓		20.37	22.93	**31.44**
✓			✓	20.79	24.93	**31.72**
	✓	✓		**40.46**	40.40	38.77
	✓		✓	40.29	**40.34**	38.86
		✓	✓	**2.81**	0.39	2.40
✓	✓	✓		**53.11**	46.39	49.23
✓	✓		✓	**52.88**	46.25	49.25
✓		✓	✓	20.54	22.79	**31.72**
	✓	✓	✓	**40.39**	**40.40**	38.86
✓	✓	✓	✓	**53.01**	46.39	49.25
mean				30.94	29.57	**32.35**

Table 3. Performances comparison of using different modalities in mIoU(%) on DeLiVER dataset. Bold values represent the best performances to the nearest rounding for each combination of modalities.

				without adaptive discounting		with adaptive discounting	
RGB	AoLP	DoLP	NIR	ECoLaF-DeepLabV3+	ECoLaF-Segformer	ECoLaF-DeepLabV3+	ECoLaF-Segformer
✓				43.48	**48.11**	43.49	46.48
	✓			5.74	2.49	**21.45**	10.45
		✓		14.42	3.29	**35.44**	19.84
			✓	7.86	2.11	**32.81**	16.79
✓	✓			43.70	**48.11**	43.36	46.48
✓		✓		44.96	**48.38**	44.95	48.11
✓			✓	44.49	**48.29**	44.39	45.01
	✓	✓		15.86	3.29	**36.35**	27.61
	✓		✓	8.05	2.11	**36.81**	13.14
		✓	✓	20.89	3.36	**41.53**	27.19
✓	✓	✓		45.61	48.38	45.26	**48.75**
✓	✓		✓	45.03	**48.29**	44.25	47.77
✓		✓	✓	45.93	48.29	45.57	**49.85**
	✓	✓	✓	21.58	3.36	**41.72**	33.31
✓	✓	✓	✓	46.20	49.28	45.74	**49.85**
mean				30.25	27.14	**40.21**	36.04

Table 4. Performances comparison of using different modalities in mIoU(%) with and without adaptative discounting on MCubeS dataset. Bold values represent the best performances to the nearest rounding for each combination of modalities.

				without adaptive discounting		with adaptive discounting	
RGB	Depth	Event	LiDAR	ECoLaF-DeepLabV3+	ECoLaF-Segformer	ECoLaF-DeepLabV3+	ECoLaF-Segformer
✓				1.13	12.67	25.20	**31.44**
	✓			41.53	**41.70**	40.35	38.77
		✓		1.13	**1.99**	0.39	1.87
			✓	1.13	1.99	0.12	**2.40**
✓	✓			45.51	**49.81**	46.25	49.23
✓		✓		1.13	12.67	22.93	**31.44**
✓			✓	1.13	12.67	24.93	**31.72**
	✓	✓		**42.69**	41.70	40.40	38.77
	✓		✓	41.60	**41.69**	40.34	38.86
		✓	✓	1.13	1.99	0.39	**2.40**
✓	✓	✓		46.41	**49.81**	46.39	49.23
✓	✓		✓	45.50	**49.81**	46.25	49.25
✓		✓	✓	1.13	12.67	22.79	**31.72**
	✓	✓	✓	**42.77**	41.69	40.40	38.86
✓	✓	✓	✓	46.42	**49.81**	46.39	49.25
mean				24.02	28.18	29.57	**32.35**

Table 5. Performances comparison of using different modalities in mIoU(%) with and without adaptative discounting on DeLiVER dataset. Bold values represent the best performances to the nearest rounding for each combination of modalities.

grant 2024-AD011014391 on the supercomputer Jean Zay's A100 partition along with computing resources of CRIANN (Normandy, France).

References

[1] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293*, 2015. 1, 2

[2] Tim Brödermann, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Hrfuser: A multi-resolution sensor fusion architecture for 2d object detection. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 4159–4166, 2023. 2

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 2, 4

[4] A.P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967. 1, 2

[5] Thierry Denoeux. A neural network classifier based on dempster-shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(2):131–150, 2000. 2, 4

[6] Lucas Deregnaucourt, Alexis Lechervy, Hind Laghmara, and Samia Ainouz. An evidential deep network based on dempster-shafer theory for large dataset. *Advances and Applications of DSmT for Information Fusion*, page 907, 2023. 1, 2, 5

[7] Jean Dezert and Florentin Smarandache. A new probabilistic transformation of belief mass assignment. *CoRR*, abs/0807.3669, 2008. 3, 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[9] 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. 5

[10] Jamil Fayyad, Mohammad A Jaradat, Dominique Gruyer, and Homayoun Najjaran. Deep learning sensor fusion for autonomous vehicle perception and localization: A review. *Sensors*, 20(15):4220, 2020. 1

[11] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Gläser, 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, 2021. 2

[12] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023. 1

[13] Z. Guo, Z. Wan, Q. Zhang, X. Zhao, Q. Zhang, L.M. Kaplan, A. Jøsang, D.H. Jeong, F. Chen, and J.-H. Cho. A survey on uncertainty reasoning and quantification in belief theory and its application to deep learning. *Information Fusion*, 101, 2024. 2

[14] Abhishek Gupta, Alagan Anpalagan, Ling Guan, and Ahmed Shaharyar Khwaja. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array*, 10:100057, 2021. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[16] Anne-Laure Jousselme, Dominic Grenier, and Éloi Bossé. A new distance between two bodies of evidence. *Information fusion*, 2(2):91–101, 2001. 3

[17] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8602–8611, 2020. 2

[18] Çağrı Kaymak and Ayşegül Uçar. A brief survey and an application of semantic image segmentation for autonomous driving. *Handbook of Deep Learning Applications*, pages 161–200, 2019. 1

[19] Yabei Li, Junge Zhang, Yanhua Cheng, Kaiqi Huang, and Tieniu Tan. Semantics-guided multi-level rgb-d feature fusion for indoor semantic segmentation. In *2017 IEEE international conference on image processing (ICIP)*, pages 1262–1266. IEEE, 2017. 2

[20] 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. 2, 5, 6, 7

[21] 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. 5

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[25] Arnaud Martin. About conflict in the theory of belief functions. In *Belief Functions: Theory and Applications: Pro-*

ceedings of the 2nd International Conference on Belief Functions, Compiègne, France 9-11 May 2012, pages 161–168. Springer, 2012. 3

[26] 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. 2, 3, 4

[27] Lukas Schneider, Manuel Jasch, Björn Fröhlich, Thomas Weber, Uwe Franke, Marc Pollefeys, and Matthias Rätsch. Multimodal neural networks: Rgb-d for semantic segmentation and object detection. In *Image Analysis: 20th Scandinavian Conference, SCIA 2017, Tromsø, Norway, June 12–14, 2017, Proceedings, Part I 20*, pages 98–109. Springer, 2017. 2

[28] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31:3179–3189, 2018. 1, 2

[29] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, 1976. 1, 2, 3

[30] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 9441–9447. IEEE, 2020. 2

[31] P. Smets. The combination of evidence in the transferable belief model. *Transactions on pattern analysis and machine intelligence*, 12(2):447–458, 1990. 3

[32] John Sudano and Lockheed Martin. Pignistic probability transforms for mixes of low- and high-probability events. *Information Fusion - INFFUS*, 01 2001. 5

[33] Yangjie Sun, Zhongliang Fu, Chuanxia Sun, Yinglei Hu, and Shengyuan Zhang. Deep multimodal fusion network for semantic segmentation using remote sensing image and lidar data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022. 2

[34] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013. 5

[35] Hans Thisanke, Chamli Deshan, Kavindu Chamith, Sachith Seneviratne, Rajith Vidanaarachchi, and Damayanthi Herath. Semantic segmentation using vision transformers: A survey. *Engineering Applications of Artificial Intelligence*, 126:106669, 2023. 1

[36] Zheng Tong, Philippe Xu, and Thierry Denoeux. An evidential classifier based on dempster-shafer theory and deep learning. *Neurocomputing*, 450:275–293, 2021. 2

[37] Zheng Tong, Philippe Xu, and Thierry Denoeux. Evidential fully convolutional network for semantic segmentation. *Applied Intelligence*, 51(9):6376–6399, 2021. 1, 2

[38] Johan Vertens, Jannik Zürn, and Wolfram Burgard. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8461–8468. IEEE, 2020. 2

[39] Fan Wang, Samia Ainouz, Chunfeng Lian, and Abdelaziz Bensrhair. Multimodality semantic segmentation based on polarization and color images. *Neurocomputing*, 253:193–200, 2017. 2

[40] 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. 4

[41] S. Xu, Y. Chen, C. Ma, and X. Yue. Deep evidential fusion network for medical image classification. *International Journal of Approximate Reasoning*, 150:188–198, 2022. 2

[42] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 2

[43] 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. 2, 5, 6, 7, 8

[44] Yuxiao Zhang, Alexander Carballo, Hanting Yang, and Kazuya Takeda. Perception and sensing for autonomous vehicles under adverse weather conditions: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:146–177, 2023. 1

[45] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:104042, 2021. 2