Fusion of Semantic Segmentation Models for Vehicle Perception Tasks

Danut-Vasile Giurgi, Jean Dezert, Thomas Josso-Laurain, Maxime Devanne, Jean-Philippe Lauffenburger

Fusion of Semantic Segmentation Models for Vehicle Perception Tasks

Dănuţ-Vasile Giurgi[a], Jean Dezert[b], Thomas Josso-Laurain[a],
Maxime Devanne[a], Jean-Philippe Lauffenburger[a]

[a]IRIMAS-UR7499, Université de Haute-Alsace, Mulhouse, France.
[b]The French Aerospace Lab, ONERA, Palaiseau, France.
Emails: vasile.giurgi@uha.fr, jean.dezert@onera.fr, thomas.josso-laurain@uha.fr,
maxime.devanne@uha.fr, jean-philippe.lauffenburger@uha.fr

***Abstract*—In self-navigation problems for autonomous vehicles, the variability of environmental conditions, complex scenes with vehicles and pedestrians, and the high-dimensional or real-time nature of tasks make segmentation challenging. Sensor fusion can representatively improve performances. Thus, this work highlights a late fusion concept used for semantic segmentation tasks in such perception systems. It is based on two approaches for merging information coming from two neural networks, one trained for camera data and one for LiDAR frames. The first approach involves fusing probabilities along with calculating partial conflicts and redistributing data. The second technique focuses on making individual decisions based on sources and fusing them later with weighted Shannon entropies. The two segmentation models are trained and evaluated on a particular KITTI semantic dataset. In the realm of multi-class segmentation tasks, the two fusion techniques are compared and evaluated with illustrative examples. Intersection over union metric and quality of decision are computed to assess the performance of each methodology.**

Keywords: segmentation models, vehicle perception, belief functions, PCR6 fusion rule, entropy.

I. Introduction

A. Perception for autonomous driving

In perception systems of self-driving cars, neural networks (NN) already serve as strong pillars for several tasks such as computer vision. Convolutional neural networks (CNNs) [4], [5], and visual transformers [6] are now more resilient and efficient in image segmentation. However, in the top of this NNs efficiency, architectures of various dimensionalities and large number of parameters can benefit from information fusion and data merging.

Thus, fusing sensors can significantly impact the performance of algorithms designed for autonomous driving tasks. Aleatoric uncertainty arises from sensors, while epistemic uncertainties occur from the models, leading to imprecise information and conflicts. Unlike aleatoric uncertainty, epistemic uncertainty can be reduced through increased data and model confidence [1], [2]. When merging data, the improvement in performance can depend on the choice of fusion technique and its implementation. In terms of fusion, late fusion (merging information at decision level) has been explored to potentially enhance the model's output by providing better reasoning on which source of information impacts the final decision more.

For self-driving tasks, various sensors like cameras, LiDARs, or radars are used to perceive the environment from different point of view. Some sensors have a greater impact on decision-making than others. Additionally, computer vision tasks like segmentation pose challenges for autonomous vehicles due to environmental variability, complex scenes, and real-time requirements [3]. A well-designed fusion strategy can significantly improve final output results.

B. Multi-modal evidential information fusions

Common approach in fusion is to convert information sources into evidential formulations using belief functions. Many studies are exploring the incorporation of Dempster-Shafer's theory into deep neural networks to achieve plausible reasoning. In self-navigation tasks, works such as [7], [8], and [9] demonstrate the integration of multi-sensor information or the combination of neural networks with evidence theory for perception tasks like object recognition (road, pedestrian, etc.) or area partitioning within segmentation.

Furthermore, models such as evidential cross-fusion approach, inspired by [10], incorporate cross-fusion alongside the neural network layer, enabling information exchange between sensors. These methodologies stand out for their ability to handle situations with imprecise data and conflicting sources by introducing a new class responsible for uncertainties. The concept revolves around using distance to prototypes for road detection or segmentation. Instead of a probabilistic approach, an evidential formulation based on belief functions is employed, and decisions rely on Dempster-Shafer's (DS) rule of combination. Once a fusion rule is applied, decision-making can range from a simple *argmax* function to more sophisticated methods (e.g. Decision Based on Interval [11]). Dempster's rule of combination can be easily extended to combine more than two sources of evidence. Its commutative and associative mathematical properties make the DS rule appealing for implementation, permitting sequential fusion, without altering the fusion results, regardless of the data merging order. While the DS rule usually produces good results for self-driving tasks, there are scenarios where its dictatorial behavior yields wrong decisions. Some alternatives and comparative analyses between DS and other fusion rules

are available in [12], [13]. Thus, while [9] presented the DS rule internally, by adding a mass-function merging module at the end of the architecture [14], this work focuses on fusing probabilities generated from the softmax of each architecture.

C. Contributions

In [10], three fusion approaches are mentioned: early fusion, cross fusion, and late fusion. Among these, the second one is preferred due to the advantage of progressively exchanging features. However, the focus of [10] is more on road detection, emphasizing neural network capabilities over fusion techniques.

Late fusion of data can preserve important features captured by sensors, impacting the final decision. Since a multi-modal system is considered with two sources, this work proposes approaching a late fusion. Thus, the focus is on more suitable rules to merge two individual models for multi-class segmentation. The first method is based on the Proportional Conflict Redistribution (PCR) rule, while the second one weights fused decisions based on their quality computed from Shannon entropy [15].

The PCR fusion has seen improvements over the years (PCR5, PCR6, and PCR6$^+$), considering the complexity, mathematical properties, sophisticated conflict management and the number of information sources. PCR5, PCR6, and PCR6$^+$ rules are not associative and they can be used to fuse more than two sources of evidence altogether. PCR5 and PC6 do not preserve the neutrality of the vacuous BBA in the fusion process, whereas PCR6$^+$ does. These three aforementioned rules coincide when we have only two sources of evidence to combine. Further details about these rules can be found in [16].

The work continues with the following sections: Background (basics about belief functions, fusion rules, and decision-making), Implementation (architecture and exemplifications of fusion rules), Results (dataset and results for proposed techniques), and Conclusion.

II. Background

A. Fusion and data representation

Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$ represent the universe of elements, known as the *frame of discernment* (FoD), with mutually exclusive elements of single cardinality referred to as *singletons* [19]. The notation of the mass function, or the *basic belief assignment* (BBA), $m(\cdot)$ denotes a distribution $m : 2^\Theta \to [0, 1]$, satisfying:

$$m(\emptyset) = 0, \text{ and } \sum_{X \subseteq \Theta} m(X) = 1. \tag{1}$$

where 2^Θ denotes the power set[1] of Θ.

Here, $m(X)$, known as the mass of element (i.e., subset) X of Θ, represents the available evidence committing to event X by the source of evidence. The definition (1) corresponds to Shafer's closed world assumption. A subset X is termed a *focal element* of $m(\cdot)$ if and only if $m(X) > 0$. With these notations from the belief theory, different fusion rules can be defined to combine several distinct (i.e. cognitively independent) sources of evidence. The fundamentals of evidence formulation can be easily translated to Bayesian theory as long as only singletons are considered and uncertainty is not taken into account.

1) PCR6$^+$ Fusion Rule: The expression for the PCR6$^+$ fusion of $S \geq 2$ BBAs can be computed using the formula $m^{\mathrm{PCR}^+}_{1,2,\ldots,S}(\emptyset) = 0$, where each $A \in 2^\Theta \setminus \emptyset$. The final computation formula is:

$$\begin{aligned} m^{\text{PCR6}^+}_{1,2,\ldots,S}(A) &= m^{\text{Conj}}_{1,2,\ldots,S}(A) \\ &+ \sum_{j \in \{1,\ldots,\mathcal{F}\} | A \in \mathbf{X}_j \wedge \pi_j(\emptyset)} \left[\left(\kappa_j(A) \sum_{i \in \{1,\ldots,S\} | X_{j_i} = A} m_i\left(X_{j_i}\right) \right) \right. \\ &\left. \cdot \frac{\pi_j(\emptyset)}{\sum_{X \in \mathbf{X}_j} \left(\kappa_j(X) \sum_{i \in \{1,\ldots,S\} | X_{j_i} = X} m_i\left(X_{j_i}\right) \right)} \right], \end{aligned} \tag{2}$$

where $m^{\text{Conj}}_{1,2,\ldots,S}(.)$ is the classical conjunctive fusion rule [16], $\mathcal{F}$ is the cardinality of the cartesian product of the sets of focal elements of the S BBAs, $\kappa_j(A)$ and $\kappa_j(X)$ denote the binary indices representing factors A and X, respectively, which contribute to conflicting aspects within the product. Detailed expressions for computing element $\pi_j(\emptyset)$ can be found in [16], [17] with Matlab™ code included.

2) Shannon Entropy: In the probabilistic context, Shannon entropy serves as a common tool for quantifying uncertainty and distributing randomness. Low entropy values indicate more precise predictions or certainty, whereas high entropy suggests uncertainty and misinformation. Shannon's entropy, denoted[2] by $H(P_N)$, is defined as:

$$H\left(P_N\right) \triangleq -\sum_{i=1}^{|\Theta|} P\left(\theta_i\right) \log\left(P\left(\theta_i\right)\right), \tag{3}$$

where, by convention, $P\left(\theta_i\right) \log\left(P\left(\theta_i\right)\right) = 0$ is taken if $P\left(\theta_i\right) = 0$. Shannon entropy is expressed in *nats* if natural logarithm function is taken in (3), or it is expressed in *bits* if the logarithm function is considered in base 2 (i.e. $\log_2$).

The maximum of Shannon entropy is obtained with the uniform pmf defined by $P^{\text{unif}}(i) = 1/N$ for $i = 1, 2, \ldots, N$, and we have

$$H^{\max} = H(P_N^{\text{unif}}) = \log(|\Theta|) = \log(N). \tag{4}$$

Consequently, assuming that some decisions can be made, these decisions are fused by weighting their quality calculated from the entropy. Once a weighting factor is computed with respect to the highest entropy (to work with normalized entropy), the decisions can be fused by a simple weighted averaging rule (See Section III-B2).

[1] 2^Θ is the set of all subsets of Θ including the empty set $\emptyset$ and Θ itself.

[2] Here P_N is a probability mass function (pmf) defined over FoD $\Theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$.

B. Decision-Making

Considering a probability mass function obtained with a fusion rule, the last step involves decision-making to obtain the results. Thus, the elements of m_{PCR6^+} represent decisions with respect to the focal elements, while $\hat{X}$ represents the final decision. The final decision is characterized by the maximum values among the fused mass functions of m_X, where $X \in 2^\Theta \setminus \{\emptyset\}$. In this work, $argmax$ is used for the final decision; therefore, $\hat{X}$ is defined as:

$$\hat{X} = \arg \max_{X \in 2^\Theta \setminus \{\emptyset\}} m_X \tag{5}$$

where the values are computed based on the applied fusion technique. Mass functions $m_X(\cdot)$ represent the obtained values through fusion for each chosen focal element X (each class, typically the road (R), the vehicles (V) and the background (B)). This step is necessary when using the PCR6+ fusion scheme, not for the entropy computation

After reaching the final decision, the quality indicator can be computed according to [11]. This evaluates the fairness of the decisions. This confidence factor, denoted as $q(\hat{X})$, is given by:

$$q(\hat{X}) \triangleq 1 - \frac{m_X}{\sum_{X \in 2^\Theta \setminus \{\emptyset\}} m_X} \tag{6}$$

The larger the value of the quality indicator, the more confidence the model has.

Once the BBAs representing the evidence in the corresponding pixels are evaluated, a final task remains: to determine the classes for each pixel. Therefore, given the previous statement (5) as presented in [11], the quality of decisions can be computed with respect to classes.

III. Implementation

A. Neural Network Architecture

For the implementation, two identical segmentation models are considered. They are both CNN-based, similar with the two pipelines of a cross-fusion model [10], but individual. In this case, one represents the neural network architecture that is trained with camera images, and another identical one is used to learn features from dense map LiDAR data.

The dense depth map images are obtained from 3D points clouds using projection and translating matrices [21]. Both of the models have 24 layers, following an architecture[3] with encoder, context module, and decoder like illustrated in Fig. 1. The model uses convolutions and up-sampling operations, regularized with Dropout. The paper itself is mainly oriented to emphasize the strength of fusion rather than how efficient the neural network models are. In this way, the effects of fusing two NN models, one based on LiDAR and the other based on camera features allow interchanging meaningful information from the two sources.

By using a common CNN architecture, the output represents probability distributions formed out of logits, as a result of the last layer. The last layers are softmax activation functions with values mapped between 0 and 1. Thus, the models represent two independent sources of information with probability outputs.

Fig. 2 shows more in detail the fusion approach block from Fig. 1. It can be observed how the two segmentation models are positioned with respect to fusion rule and decision making.

[3]https://github.com/vasigiurgi/fusing-segmentation-models

B. Fusion of the segmentation models

As mentioned, once the neural networks are computed, two approaches are intended to be applied from a fusion perspective. Following, the two sources of information have 3 classes each: R (Road), V (Vehicle), and B (Background).

1) PCR6$^+$ fusion: Below, the algorithm for the PCR6$^+$ is briefly explained, see [16] or [17] for details.

Algorithm 1: PCR6$^+$ fusion

Input:
$m_L = (m_1(R), m_1(V), m_1(B))$
$m_C = (m_2(R), m_2(V), m_2(B))$
Output: mPCR6plus: m_1 fused with m_2

```
NbrSources ← sources of information;
CardTheta ← calculate cardinality (FoD);
Combinations ← generate all combinations of
               sources;
for c in Combinations do
    PC ← current combination;
    massConj ← calculate mass conjunction for PC;
    Intersections ← calculate intersection of sources;
    if Intersections ≠ 0 then
        update mPCR6plus based on PC;
    end
    else
        calculate contributions from each source;
        update mPCR6plus;
    end
end
return mPCR6plus;
```

For instance, suppose that the (Bayesian) BBA for the camera is:

$$[m_1(R) = 0.8, \quad m_1(V) = 0.15, \quad m_1(B) = 0.05]$$

and for the LiDAR:

$$[m_2(R) = 0.55, \quad m_2(V) = 0.25, \quad m_2(B) = 0.20]$$

The result of masses for PCR6$^+$ (m_{PCR6^+}, noted m_f) fusion will output:

$$[m_f(R) = 0.845, \quad m_f(V) = 0.145, \quad m_f(B) = 0.010]$$

resulting in the final decision on the singleton R which represents the road.

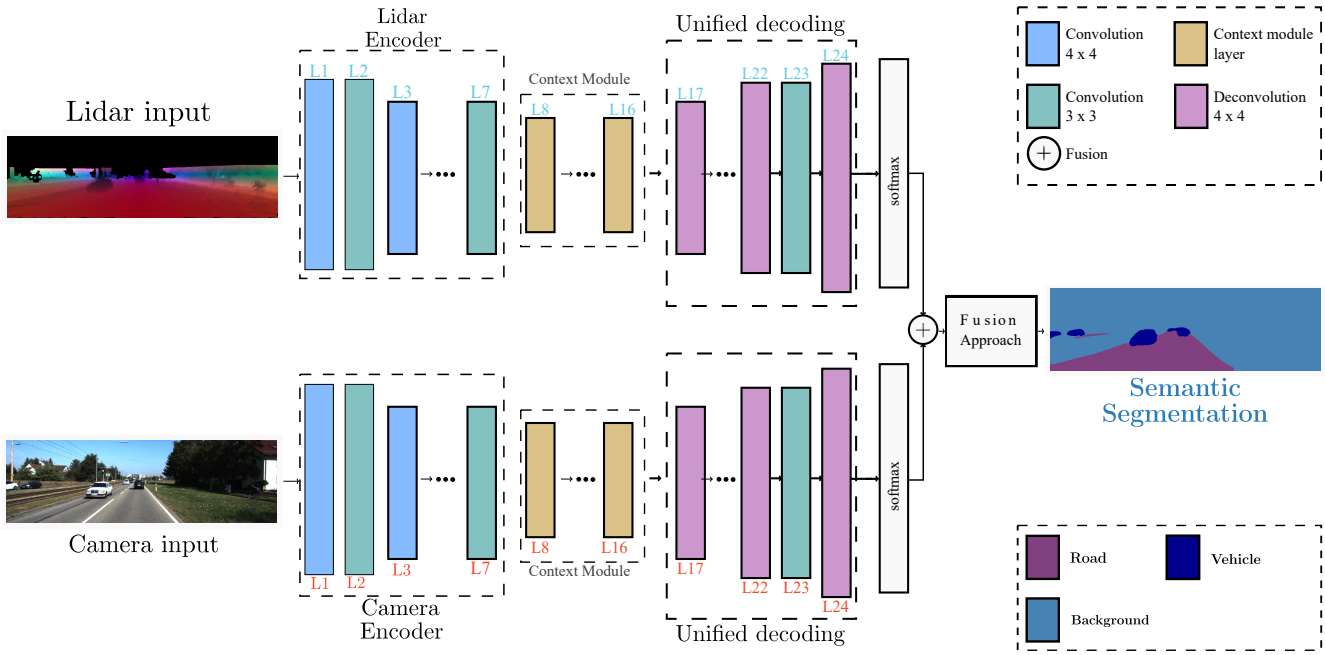


Figure 1: Fusion of segmentation models: LiDAR model and camera model.

2) Weighting quality decisions fusion based on Shannon's Entropy: The second approach works by making some decisions based on the Bayesian output of the architectures considering entropies thereafter to check how consistent information is. Suppose that for a camera model, a pixel (i, j), is considered with the following mass values for each class:

$$[m_1(R) = 0.80, \quad m_1(V) = 0.15, \quad m_1(B) = 0.05]$$

In this situation, taking the decision for pixel $(i, j) = R$ (from camera model) can be relevant, but not 100% sure because $m_1(R) < 1$. In the same way, for a LiDAR frame, suppose a pixel with mass values accordingly:

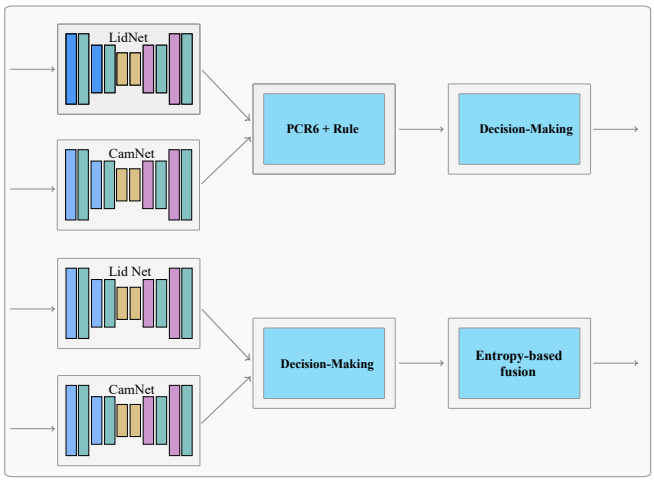


Figure 2: Fusion approach diagram.

$$[m_2(R) = 0.55, \quad m_2(V) = 0.25, \quad m_2(B) = 0.20]$$

The decision will be the same, pixel $(i, j) = R$, which can be again relevant, but the decision tends to be riskier as $m_2(R)$ is just above 0.5. Instead of fusing directly probabilities, another way is to fuse weighted decisions by their quality calculated from entropy. Consequently, in the previous example, based on m_1, the early state of the decision will represent R (road) class for the camera segmentation model:

$$[md_1(R) = 1, \quad md_1(V) = 0, \quad md_1(B) = 0].$$

Then accordingly to the weight, the decision will be updated. The weight of source 1 for this pixel is calculated by the quality measure as:

$$w_1 = 1 - \frac{H(m_1)}{H^{\max}},$$

where $H(m_1)$ is the entropy of m_1 because m_1 is Bayesian[4]. Therefore, $H(m_1)$ corresponds to Shannon entropy, while $H^{\max}$ is the maximum of Shannon entropy obtained for a uniform probability mass function as highlighted in Section 3.

Based on m_2, the R class will be decided, therefore judgment based on LiDAR data is:

$$md_2(R) = 1, \quad md_2(V) = 0, \quad \text{and } md_2(B) = 0,$$

[4]For a more general (non-probabilistic) context when working with non-Bayesian BBAs we could use the generalized entropy for belief function defined in [18].

with the weight of source 2 (LiDAR) provided by the quality:

$$w_2 = 1 - \frac{H(m_2)}{H^{\max}}.$$

Then the decisions are fused by a simple weighted averaging rule as follows:

$$md(R) = \frac{w_1}{w_1 + w_2} \cdot md_1(R) + \frac{w_2}{w_1 + w_2} \cdot md_2(R),$$
$$md(V) = \frac{w_1}{w_1 + w_2} \cdot md_1(V) + \frac{w_2}{w_1 + w_2} \cdot md_2(V),$$
$$md(B) = \frac{w_1}{w_1 + w_2} \cdot md_1(B) + \frac{w_2}{w_1 + w_2} \cdot md_2(B).$$

In this simple example $|\Theta| = 3$ since the FoD has 3 singletons only (Eq. 4, where N = 3). Therefore, w_1 will have a greater value than w_2 due to the lower entropy of $H(m_1)$. Consequently, the camera source shows greater confidence.

IV. Experimental Results

A. LiDAR-Camera Dataset and Networks Training

The LiDAR-camera dataset, known as the semantic KITTI dataset, initially comprises 200 camera images. This dataset closely resembles the structure of the KITTI Stereo and KITTI Flow 2012/2015 datasets. However, officially, the KITTI semantic dataset includes no LiDAR frames, such as those found in the road dataset. Henceforth, this particular KITTI semantic adds the lidar point clouds. Further, the 3D point-cloud points corresponding to the existing camera frames need to be identified within the extensive original KITTI raw dataset [20], which encompasses data for all tasks.

To augment the dataset, LiDAR frames have been successfully projected and up-sampled for 127 out of the 200 camera images, resulting in dense depth images. The mapping of a 3D LiDAR point $\boldsymbol{x}$ to a point $\boldsymbol{y}$ in the camera plane is accomplished through the application of the KITTI projection $\boldsymbol{P}$, rectification $\boldsymbol{R}$, and translation $\boldsymbol{T}$ matrices [21].

$$y = PRT\ x \tag{7}$$

To address the sparsity of the projected LiDAR scan, an up-sampling technique is applied to generate a dense depth map, as illustrated bellow, in Fig. 3. The up-sampling process follows the methodology detailed in [9] and [21].

The LiDAR frames presented earlier are integrated alongside the camera ones in the two segmentation models, each individually with the same ground truth for both pipelines. The masks are represented by three elements corresponding to the classes: road (magenta), vehicle (dark blue), and background (blue), according to the original annotation as illustrated in Fig 4b.

Camera and LiDAR models receive 127 input frames, partitioned into 114 for training and 13 images for validation. Each architecture is trained individually. The segmentation models are trained for 50 epochs. For the hyperparameters, mean-squared error and Adam optimizer are used.

B. Segmentation Performance Analysis

For the performance assessment, the intersection-over-union metric is considered accordingly to [22], and the formula (8).

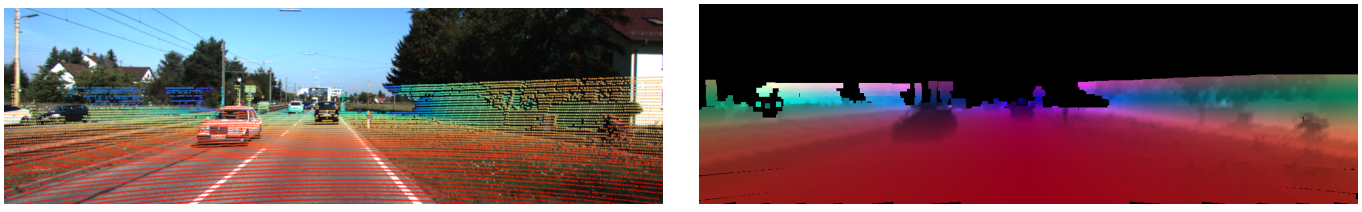
$$IoU = \frac{TP}{TP + FP + FN} \tag{8}$$

where TP, FP, and FN represent the true positives, false positives, and false negatives, respectively. Since the models are fused using two methods, the metric evaluations will be computed for each method separately, namely PCR6$^+$ fusion and fused decisions with entropy and compared at the end. Initially, the IoU is calculated individually for each frame, while at the end the average IoU, noted IoU is presented for results evaluation.

Table I: Performance Metrics.

	Fusion schemes	
Metric	**PCR6$^+$**	**Entropy**
Mean IoU	0.7626	**0.8009**
Highest IoU	0.8780 (img 9)	0.8954 (img 7)
Lowest IoU	0.6074 (img 1)	0.6540 (img 1)
Class-wise IoU		
Class R	0.7750	0.8056
Class V	0.5909	0.6590
Class B	0.9219	0.9382

The performances with respect to the IoU metric are shown in Table I. It can be observed that both approaches share the worst case (image 1), but for the best value of IoU, the fusion method impacts the output differently. Thus, on average, the decision fusion method with the entropy performs better ($IoU = 0.8009$) than the PCR6$^+$ ($IoU = 0.7626$). The IoU metric is also calculated for each class individually. As expected, the class representing vehicles performs the worst

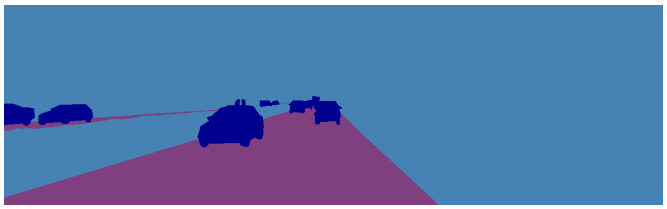


(a) Projection of LIDAR over the camera image

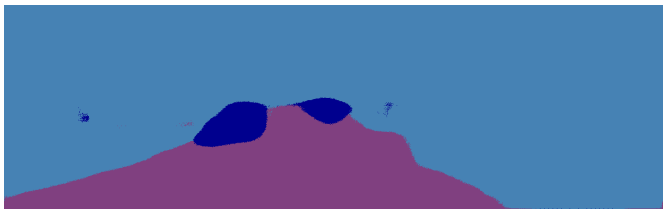
(b) Dense depth map image from LiDAR point clouds

Figure 3: LiDAR point clouds pre-processed to 2D images.

(a) Input frame for camera segmentation model



(b) Simplified ground truth (R, V and B)



(c) Prediction from the LiDAR Model

(d) Prediction from the camera model

Figure 4: (a) Train image, (b) Ground truth mask
Predictions from segmentation models: (c) LiDAR and (d) camera.

in terms of segmentation, while the class of the background is best identified. This happens because of the segmented area of the pixels related to a class. Such behavior is already inherent in what the model learns, while the fusion methods only tend to improve the quality of information provided by the source.

The set of frames, Fig. 4a and Fig. 4b represent the original image and mask, respectively Fig. 4c and Fig. 4d stand for predictions from the segmentation models of each pipeline.

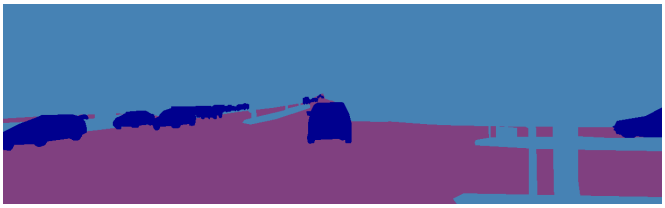
The simplified ground truth has only 3 classes (Road, Vehicle and B ackground). Since the LiDAR data is up-sampled and mapped into a 2D dense map image, some of the features are less pronounced than in the camera's case. However, there are scenarios when outputted images via LiDAR show important characteristics of the classes.

In Fig. 5, the images with the worst IoU score are highlighted within the two fusion approaches. At the top, the ground truth is provided for reference, followed by images obtained through fusion via PCR6$^+$ (Fig. 5b) and fused decisions weighted by entropies (Fig. 5c).

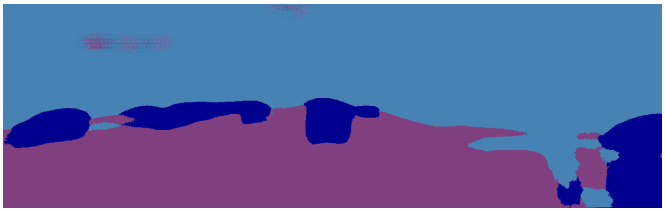
Finally, Fig. 6 presents the best case for PCR6$^+$, while Fig. 7 presents the best case for fused entropies. Here, the two fusion approaches do not share the same frame for the best case. The images from the top part of both illustrations represent the ground truth images (Fig. 6a and Fig. 7a).

The best results with respect to intersection over union are shown in the bottom part of the illustrations. The fusion based on PCR6$^+$ works better for Fig. 6b (image 9), while the fusion with Shannon entropies performs better for Fig. 7b (image 7 according to the table).

Considering Table I, the approach based on fused decisions demonstrates superior performance, as indicated by the higher IoU score compared to the PCR6$^+$ method. Although this example shows a performance advantage for the fused decision approach, it is important to note that PCR6$^+$ is specifically designed for multi-source information systems. The conflict



(a) Ground truth for the worst case scenario



(b) Worst IoU score for the approach-based PCR6$^+$

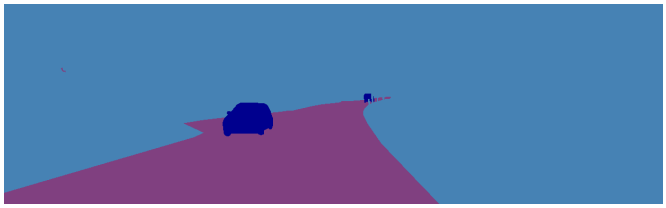
(c) Worst IoU score for
the approach-based Shannon entropy

Figure 5: Worst case scenario.

redistribution feature of PCR6$^+$ could offer significant advantages in scenarios involving a higher number of sources, potentially leading to more compelling results in such contexts.

C. Quality of Decision Assessment

The quality of the decision is calculated for the best case scenario of the entropy method, i.e. image 7, and PCR6$^+$



(a) Ground truth (img 9)

(b) Best IoU score PCR6$^+$ (img 9)

Figure 6: Best case scenario PCR6$^+$.

(a) Ground truth (img 7)

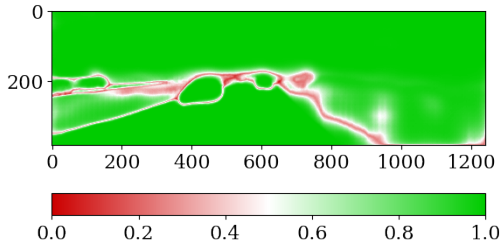
(b) Best IoU score Entropy (img 7)

Figure 7: Best case scenario Shannon entropy.

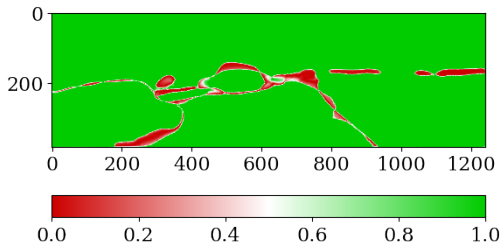
fusion, i.e. image 9. On the top part, Fig. 8a expresses how confident decisions are after a PCR6$^+$ fusion, while the bottom image, Fig. 8b shows the quality of judgments when considering fused decisions with Shannon weighted entropies (image 7).

Both illustrations show areas at the borders between classes that are more challenging for segmentation, while large areas such as background are much easier to segment efficiently. Computing a numerical percentage of pixel-related decisions over a threshold in terms of confidence results distinguishes the two methods noticeably.

Thus, the quality of decisions for fused judgments with Shannon entropy is more consistent. Regardless of the chosen threshold, whether it's 0.5 or 0.95, approximately 94.85% of decisions are confident. In contrast, judgments based on PCR6$^+$ fusion exhibit some variability. Closer to 1, e.g. threshold equals to 0.95, the decisions are less confident with a score of 75.48%, while with a threshold of 0.5, the



(a) Quality decision based on PCR6$^+$ (img 9)



(b) Quality decision based on Entropy (img 7)

Figure 8: Quality of decision for best case scenarios.

decisions seem to be more confident showing a score of 99.36%. However, this confidence can be problematic as it is very close to 0.5. This explains the gradient colors for decisions based on PCR6$^+$ fusion, and why the bottom image is less chromatically diverse.

V. Conclusion

This paper shows two ways of late fusion for two deep-learning segmentation models based on camera and LiDAR frames. The architectures are trained initially as probability models and later are fused via a conflict redistribution fusion (PCR6$^+$) or based on fused decisions with respect to Shannon entropy.

The work focuses more on the fusion techniques and their application for perceptive tasks and less on the performances of the neural networks and complexity for real-time tasks. For simplicity, three classes are considered and the approaches are assessed using mean intersection over union metric.

The method based on weighting fused decisions gives better results, but PCR6$^+$ is explicitly designed for multi-source information systems, where its conflict redistribution capability can provide substantial benefits in situations involving a greater number of sources. The results are presented from a qualitative point of view, but robustness and generalization are aimed to be considered for future investigations, as well as model optimization. In the fusion part, contextual discounting to investigate the contribution of each source is under examination as well as the fusion technique positioning.

Acknowledgments

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

REFERENCES

[1] S. Manchingal, F. Cuzzolin, *Epistemic Deep Learning*, 2022.

[2] A. Amini, W. Schwarting, A. Soleimany, D. Rus, *Deep evidential regression*,Advances In Neural Information Processing Systems, 33 pp. 14927-14937, 2020.

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

[4] A. Krizhevsky, I. Sutskever, G.E. Hinton, *Imagenet classification with deep convolutional neural networks*, Communications of the ACM, Vol. 60 (6), pp. 84–90, 2017.

[5] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, *You only look once: Unified, real-time object detection*, in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 779–788, 2016.

[6] A. Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2021.

[7] P. Xu et al., *Multimodal information fusion for urban scene understanding*, Machine Vision and Applications, Vol. 27(3), pp. 331–349, 2016.

[8] E. Capellier et al., *A deep evidential learning for arbitrary lidar object classification in the context of autonomous driving*, in Proc. of 2019 IEEE Intelligent Vehicles Symposium (IV), pp. 1304–1311, 2019.

[9] M. N. Geletu, D-V. Giurgi, T. Josso-Laurain, M. Devanne, M. M. Wogari and J-P. Lauffenburger., *Evidential deep learning-based multi-modal environment perception for intelligent vehicles*, Proc. of 2023 IEEE Intelligent Vehicles Symposium (IV), pp. 1–6, 2023.

[10] L. Caltagirone, M. Bellone, L. Svensson, M. Wahde, *LIDAR-camera fusion for road detection using fully convolutional neural networks*, Robotics and Autonomous Systems, Vol. 111, pp. 125–131, 2019.

[11] J. Dezert, D. Han, J.-M. Tacnet, S. Carladous, Y. Yang, *Decision-making with belief interval distance*, in Proc. of Belief 2016 Int. Conf., Prague, CZ, September 21–23, 2016.

[12] J. Dezert, P. Wang, A. Tchamova, *On The Validity of Dempster-Shafer Theory*, Proc. of Fusion 2012, Int. Conf. on Information Fusion, Singapore, July 2012.

[13] A. Tchamova, J. Dezert, *On the Behavior of Dempster's Rule of Combination and the Foundations of Dempster-Shafer Theory*, Proc. of IEEE IS'2012, Sofia, Bulgaria, Sept. 6–8, 2012.

[14] Z. Tong, P. Xu, T. Denœux, *Fusion of evidential CNN classifiers for image classification*, 2021.

[15] C. Shannon, W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press,1949.

[16] T. Dezert, J. Dezert, F. Smarandache, *Improvement of Proportional Conflict Redistribution Rules of Combination of Basic Belief Assignments*, Journal of Advances in Information Fusion, 2021.

[17] F. Smarandache, J. Dezert, A. Tchamova A. (Editors), *Advances and applications of DSmT for information Fusion (Collected works)*, Vol. 5, Biblio Publishing, OH, USA, December 2023.

[18] J. Dezert, *An effective measure of uncertainty of basic belief assignments*, Proc. of Fusion 2022, Int. Conf. on Information Fusion, Linköping, Sweden, July 4–7, 2022.

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

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

[21] C. Premebida et al., *Pedestrian detection combining rgb and dense lidar data*, in Proc. of 2014 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 4112–4117, 2014.

[22] M. Everingham et al., *The pascal visual object classes challenge: A retrospective*, Int. journal of computer vision, Vol. 111(1), pp. 98–136, 2015.