

MEFA: Multimodal Image Early Fusion with Attention Module for Pedestrian and Vehicle Detection

Yoann Dupas^{1,2}^a, Olivier Hotel²^b, Grégoire Lefebvre²^c and Christophe Cérin^{1,3}^d

¹*Datamove, Inria, LIG, University Grenoble-Alpes, France*

²*Orange Innovation, Meylan, France*

³*LIPN, University Sorbonne Paris Nord, France*

Keywords: Image Fusion, Multimodal Fusion, Early-Fusion, Attention, Pedestrian and Vehicle Detection, Adverse Weather.

Abstract: Pedestrian and vehicle detection represents a significant challenge in autonomous driving, particularly in adverse weather conditions. Multimodal image fusion addresses this challenge. This paper proposes a new early-fusion attention-based approach from visible, infrared, and LiDAR images, designated as MEFA (Multimodal image Early Fusion with Attention). In this study, we compare our MEFA proposal with a channel-wise concatenation early-fusion approach. When coupled with YOLOv8 or RT-DETRv1 for pedestrian and vehicle detection, our contribution is promising in adverse weather conditions (i.e. rainy days or foggy nights). Furthermore, our MEFA proposal demonstrated superior mAP accuracy on the DENSE dataset.

1 INTRODUCTION


Deep learning fusion techniques are significantly impacting a number of fields, including autonomous driving (AD) and autonomous driver assistance systems (ADAS). In particular, they address the challenge of perceiving the world and the challenge of decision-making.


World perception systems extract essential information from raw image data for decision making. They include three tasks: localization, detection, and tracking (Martínez-Díaz and Soriguera, 2018). Effective performance requires accuracy, weather robustness, efficiency with imprecise sensors, real-time processing (Xiang et al., 2023), and reduced energy consumption (Malawade et al., 2022).


However, such systems face many challenges, including the variability of object shapes, potential occlusions, variations in lighting, and the prevalence of adverse weather conditions (Martínez-Díaz and Soriguera, 2018). The process of image fusion serves to address the limitations of perception and decision-making systems. The fusion of multiple cameras allows the acquisition of data that is both rich and high-dimensional, and which is also complementary by na-


ture (Xiang et al., 2023). In (Huang et al., 2022), the authors present a taxonomy of three fusion strategies. The first of these is early fusion, or data-data fusion, which involves the merging of data that has been prepared to be spatially homogeneous. This strategy identifies correlations between image channels and generates a global fused image that is compatible with some existing object recognition models (Stahlschmidt et al., 2022), provided that the first layer of the model can be adapted to this kind of fused image. The second fusion strategy, known as deep fusion or feature-feature fusion, involves the merging of latent space features generated by a backbone module into a common latent space. Finally, late fusion, or results-results fusion, represents a further approach whereby final object detection is mixed with different strategies, including voting methods, stacking methods, and those based on Dempster-Shafer or Possibility Theory (Chen et al., 2022).

This article concentrates on the early fusion strategy, which we consider to be a valuable approach for generating fused images that include all the information required by a single-modality object detection model to perform object detection accurately. This strategy makes it possible to utilize the most recent single-modality models from the literature, such as YOLO (You Only Look Once (Terven et al., 2023)) or RT-DETR (Real-Time Detection Transformer (Zhao et al., 2024)), and to benefit from their enhanced performance. Furthermore, in our view, the use of atten-

^a <https://orcid.org/0009-0001-2080-3842>

^b <https://orcid.org/0009-0005-9688-4326>

^c <https://orcid.org/0000-0002-1325-3010>

^d <https://orcid.org/0000-0003-0993-9826>

tion mechanisms can help to improve accuracy and robustness in the context of challenging weather conditions. By assigning weights to information, it enables the neural network to identify and select relevant information from each modality and to correlate them with information from the other modality.

The paper is structured as follows. Section 2 proposes the main related and recent work on image attention-based fusion for pedestrian and vehicle detection. Section 3 explains in detail the proposed MEFA module. Section 4 presents the experimental protocol. Section 5 presents the analysis of the performances in global conditions and different weather conditions. Section 6 discusses the contributions and the results. Finally, Section 7 concludes this article and gives some perspectives.

2 RELATED WORKS

Recent studies address the topics of multimodal fusion with deep learning techniques based on attention modules (Chaturvedi et al., 2022), (Tabassum and El-Sharkawy, 2024).

The article (Chaturvedi et al., 2022) presents a deep fusion approach with its Global-Local Attention (GLA) framework aiming at improving object detection in adverse weather conditions, such as light fog, dense fog, and snow. The GLA framework utilizes multimodal sensor fusion, integrating data from cameras, gated cameras, and LiDAR at two stages: early-stage fusion through a Local Attention Network and late-stage fusion via a Global Attention Network. This dual approach allows the system to adaptively focus on the most effective sensor data based on the specific weather conditions. The GLA framework’s architecture enables it to extract local and global features, addressing the shortcomings of existing methods that typically rely on simple concatenation or element-wise addition for sensor fusion. By employing attention mechanisms, the GLA framework can dynamically allocate higher weights to the modality that exhibits better detection capabilities, thus enhancing the robustness of object detection.

In (Tabassum and El-Sharkawy, 2024), the authors introduce a multi-head attention approach to enhance vehicle detection in adverse weather conditions, specifically focusing on the MVDNet (Multimodal Vehicle Detection Network). This model integrates a multi-head attention layer to improve the processing and fusion of multimodal sensor data, such as LiDAR and radar. By employing a multi-head attention mechanism, the MVDNet can dynamically focus on various aspects of the input data, allowing for

a more comprehensive analysis and improved detection accuracy. The methodology involves two main stages: the Region Proposal Network (RPN) for generating initial proposals from sensor data, and the Region Fusion Network (RFN) for integrating these proposals. The multi-head attention layer is strategically placed within the RFN to enhance feature extraction from LiDAR and radar inputs. The paper demonstrates that the multi-head MVDNet significantly outperforms baseline models and other sensor fusion techniques.

These two papers propose some advantages. Notably, both emphasize the importance of multimodal sensor fusion to improve object detection performance in challenging weather conditions. The proposed frameworks leverage advanced attention mechanisms to dynamically adjust the focus on different sensor modalities, leading to significantly higher detection accuracy compared to traditional methods that rely on single sensors.

Nevertheless, some limitations persist. Both frameworks are intimately related to the object detection framework they use. For example, in paper (Tabassum and El-Sharkawy, 2024) they use a two-stage object detection model, which may not be adaptable to a one-stage detection model, which are more suitable for a real-time object detection. In paper (Chaturvedi et al., 2022), the GLA framework may be difficult to adapt to a new state-of-the-art model that uses a transformer-based approach, such as the ViT (Dosovitskiy et al., 2020) model.

To limit these drawbacks, we propose the MEFA module, described in the next section.

3 MEFA : MULTIMODAL EARLY FUSION WITH ATTENTION

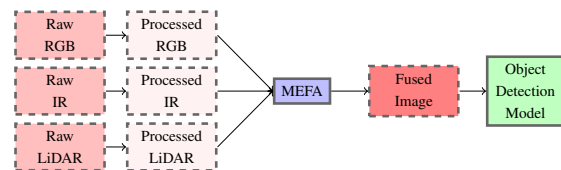


Figure 1: Overview of MEFA framework. The data obtained from the various sensors are transformed and subsequently transmitted to the MEFA module. This module generates an intermediate fused image, which represents the fusion of all input sensor data. This intermediate fused image is then provided to a single-modality object detection model, which can identify objects.

The pedestrian and vehicle detection framework, as illustrated in Figure 1, consists of multiple modules. Initially, the input images pass through image pro-

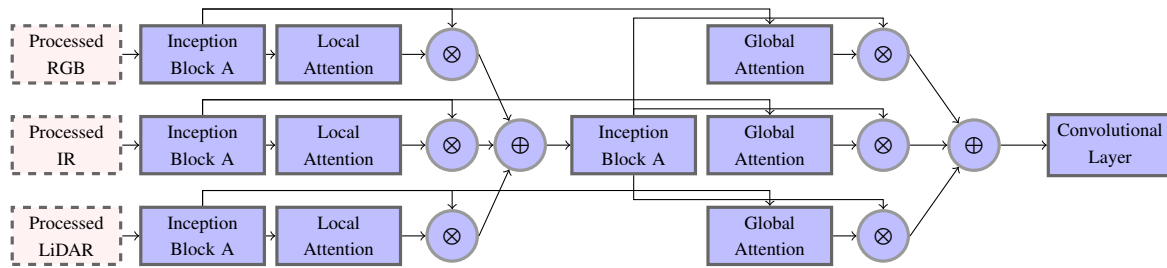


Figure 2: The MEFA module consists of three streams, where the input is initially convoluted by an inception block and subsequently provided to the local attention block. The output features are then multiplied by the output features of the inception block. The output from each stream is then combined through an addition operation. This fused output is subsequently passed to an additional inception block, after which it is sent to a global attention block. This block takes input from the first inception block for its corresponding modality. The output is then multiplied by the fused output and added, forming the final fused output. In the final step, the output is sent to a convolution layer, where it is transformed into a 3-channel output. The \otimes corresponds to the multiplication operation and \oplus corresponds to the addition operation.

cessing techniques to prepare them for fusion. Subsequently, an intermediate fused image is generated by the MEFA module, which is compatible with state-of-the-art image object detectors such as YOLOv8 or RT-DETRv1 to predict pedestrian and vehicle objects.

3.1 Input Image Processing

The data transmitted by the vehicle sensors must be transformed into a uniform image format. The DENSE dataset (Bijelic et al., 2020) used in the experiment provides data from three types of sensors: an optical camera, an infrared camera, and a LiDAR sensor. The following subsection will present a transformation operation from raw to usable data.

3.1.1 Visible and Infrared Data Processing

The optical camera provides three-channel images representing the red, blue, and green colors. In the context of working with the YOLOv8 and RT-DETRv1 models, which are dedicated to images, the transformation operations consist of scaling to a resolution of 640×640 pixels (i.e., the size of the input tensors), centering, and padding to fill in missing pixels.

The NIR (Near-InfraRed) gated camera captures photons from a specified distance by opening and closing the camera with a specified delay after emitting a pulse of near-infrared light (Grauer, 2014). It provides a batch of three grayscale images (i.e., one channel) with a time difference of 100ms between each image. The three images have been merged into a single composite image, in order to align with the input specifications of the YOLOv8 and RT-DETRv1 models.

3.1.2 LiDAR Data Processing

The LiDAR sensor provides raw data in the form of a point cloud. A point cloud can be represented in three ways: point-based, voxel-based, or 2D mapping-based (Huang et al., 2022). The 2D mapping or view-based representation is constructed by projecting points onto the camera coordinate system. This approach allows alignment with other sensors and enables the direct utilization of 2D convolutional neural network architectures. To ensure data homogeneity, a camera plane map representation on the optical camera coordinate was selected. This encoding provides an image where the points are aligned with the optical camera image.

3.2 MEFA Module

As illustrated in Figure 2, we used approaches similar to those utilized in deep fusion. In the first step, each image stream is processed through convolution layers, followed by attention layers. In the second step, the feature maps from each stream are fused before passing through a new stage of convolution layers and attention layers. The output goes through another stage of convolution layers to parameterize the channel dimension of the output.

In contrast to deep-fusion approaches, convolution layers are not designed to create a high-level representation feature space; rather, they are intended to generate a new image representation containing all relevant features from each modality. The attention layer serves as a guide to select the specific information required by the object detection model. In the initial stage, it selects information from each stream, regardless of the other modalities. In the subsequent stage, a second attention layer is designed to utilize the information from each modality to filter the fused information from the preceding stage, thereby identi-

fy only the relevant features.

The implementation of the MEFA module uses the Inception Block version A from the Inception V3 model (Szegedy et al., 2016) as convolution layers and the Global Local Attention framework (Chaturvedi et al., 2022) as attention layers. The local attention network is used as the initial attention layer for each stream. The global attention network is used as the second attention stage with the fused feature map of each stream. The fusion operation is the concatenation. The final convolution operation acts as a channel operator, parameterizing the channel output dimension to a three-channel image.

3.3 Object Detection Models

The MEFA module provides a fused image as an output, which can be used with any single-modality object detection model. Here we look at two main models from the literature.

3.3.1 YOLOv8

YOLOv8 (Jocher et al., 2023) is part of the lineage of YOLO (You Only Look Once) object detection models. The end-to-end single-shot detector architecture offers a significant advantage for real-time applications, representing a state-of-the-art model in terms of speed and accuracy.

The model has three main parts: the backbone extracts image features; the neck fuses these features; and the head predicts bounding box coordinates, object presence scores, and classification probabilities. The model employs the Complete Intersection over Union (CIoU) loss and the Distribution Focal loss (DFL) during training. This approach enables enhanced performance, particularly in the case of smaller objects (Terven et al., 2023). YOLOv8 is an anchor-free model, whereby means that the output is bounding box coordinates rather than offsets from existing anchors. The backbone is a modified CSPDarknet53 backbone with a new C2f module (faster cross-stage partial bottleneck with two convolutions). This module allows high-level features to be combined with contextual information (Terven et al., 2023).

3.3.2 RT-DETRv1

RT-DETRv1 (Zhao et al., 2024) is a hybrid object detection model using a convolutional neural network in conjunction with Transformers layers. The end-to-end NMS-free (Non-Maximum Suppression) architecture model consists of four distinct components: a backbone, a hybrid coder with Attention-

based Intra-scale Feature Interaction (AIFI), a CNN-based Cross-scale Feature Fusion (CCFF) layer, and a decoder with an uncertainty-minimum query selection scheme. The function of the backbone is to extract features from the input image. The hybrid encoder uses these one-stage CNN features to generate encoder features with the AIFI module. The CCFF module merges the multi-scale features into a feature map for the decoder. The final stage is the decoder, which uses the output of the hybrid encoder to predict coordinates and object class. The uncertainty-minimal query selection allows to optimize of the query output of the encoder to select higher quality features for the decoder head to predict the class and localization of the object.

4 EXPERIMENTAL SETUP

The experiment was conducted utilizing the DENSE database (Bijelic et al., 2020). The database comprises approximately 13,000 images of driving scenes captured under a variety of light and weather conditions, including day and night, clear, snow, light, and dense fog. The images were captured using an optical camera, a NIR gated camera, and a LiDAR sensor. Two datasets were prepared for training and testing purposes, with joint annotation from the three modalities. The training set contains approximately 100,000 objects, while the test set contains approximately 20,000 objects. The objects included in the datasets are of two main types: pedestrian and vehicle. The vehicle category includes rideable vehicles, large vehicles, vehicles, and passenger cars.

In this study, the proposed MEFA module is evaluated by comparing it with an early fusion method, i.e., channel-wise concatenation. The MEFA module is evaluated on the MEFA 3c version, which outputs an image with three channels. Each approach is evaluated with two object detection models, namely the YOLOv8 and the RT-DETRv1 models. For the RT-DETRv1 model, we employed the pre-trained weights on the COCO dataset.

To prepare the early fusion data, we combined data from the visible, infrared, and LiDAR modalities at the channel level. The labels used for the ground truth were labels from all modalities in the visible camera coordinate systems. The input layer of the YOLOv8 and RT-DETRv1 models was adapted to support a tensor of nine channels.

In order to ensure a fair and accurate benchmark, we established a well-defined experimental protocol as follows. The initial stage of the process is to identify the optimal hyperparameters for the object detec-

tion model. The search is conducted on the model that has been adapted to the channel-wise concatenation, and the optimal hyperparameters are employed for all of the compared approaches concerning the object detection model. The hyperparameters are optimized through a grid search on one single fold of the five k-folds of the training dataset. The hyperparameters include the optimizer, the weight of the box loss, the weight of the classification loss in the total loss function, the initial and final learning rate, the optimizer with momentum factor, the number of epochs for learning rate warm-up with initial momentum, the L2 regularization term, and the hyperparameters for image augmentation techniques. Once these hyperparameters are identified, a final model is trained for one hundred epochs to reach the performance plateau.

The hyperparameters search indicates the use of the SGD optimizer, with a learning rate of 0.005 for YOLOv8 and 0.008 for RT-DETRv1, as well as weight of the box loss of 0.03 and weight of the classification loss of 0.66 for YOLOv8 and 0.18 and 0.57 for RT-DETRv1. With regard to the MEFA module, the decision was taken to fix the Inception A pool feature at 32, the local and global attention intermediate output and final output feature at 64 and 256, respectively.

5 EXPERIMENTAL RESULTS

In this section, we first present the quantitative results with global object accuracy on the DENSE dataset and for each weather condition. Secondly, the qualitative results are presented to offer a more detailed explanation of the advantages and limitations of the proposed approach.

5.1 Object Detection Accuracy

5.1.1 Overall Performances

Table 1 shows the mean average accuracy with IoU at 50% (mAP_{50}) on the validation and test sets with the two object detection models (YOLOv8 and RT-DETRv1). It can be observed that when the MEFA module is combined with YOLOv8, the performance improves in comparison to the channel-wise concatenation approach. These improvements are on average 0.85% across the five folds and up to 1% on the test set. The MEFA module, when combined with RT-DETRv1, also outperforms channel-wise concatenation by 1.13% on the five folds and up to 0.5% on the test set. This observation demonstrates the benefits of using the MEFA module to fuse and select relevant

information from multimodal images.

Table 1: Accuracy results in mAP_{50} on the DENSE dataset.

Models	Mean accuracy (Validation)	Top 1 (Validation)	Accuracy (Test)
Channel-Concatenation + YOLOv8	69.03% \pm 0.64	69.65 %	70.80 %
MEFA + YOLOv8	69.88% \pm 0.70	70.78 %	71.80%
Channel-Concatenation + RT-DETRv1	72.55% \pm 0.65	73.12 %	73.60%
MEFA + RT-DETRv1	73.75% \pm 0.53	74.42 %	74.30%

5.1.2 Performances According to Weather Conditions

Figure 3 illustrates the recall values of MEFA and the channel-wise concatenation approach with YOLOv8 and RT-DETRv1 according to the weather and time of day on the test set. Images captured in clear and snowy weather represent 90% of the total images (62.7% and 27.4%, respectively). Labels in light fog, heavy fog, and rain represent 6.1%, 3%, and 0.7% of the images, respectively. The recall score was calculated on the prediction with a confidence score of 0.5 or above, and a detection was considered positive if the intersection over the union was 0.5 or above.

In contrast to channel-wise concatenation, the integration of MEFA with YOLOv8 yields an increase in object detection of 1.80%, 2.91%, 1.20%, 2.78%, and 1.33% for clear, dense fog, light fog, rain, and snow, respectively, across both daytime and nighttime conditions. MEFA, in combination with the RT-DETRv1 model, has been observed to enhance object detection by 3.94%, 3.39%, 3.67%, 2.78% and 4.04%, respectively, when compared to clear, dense fog, light fog, rain and snow conditions, taken together on a daytime and night-time basis. In general, MEFA has been found to improve recall accuracy across all weather conditions.

5.2 Qualitative Results

5.2.1 Intermediate Features Visualization

Figure 4 depicts the intermediate fused image of the MEFA module output. It can be observed that the visible features represent most of the resulting images. In this process, we analyze feature maps from local attention outputs during the inference. In these outputs, only a small number of features are retained during local attention for the infrared and LiDAR input, whereas almost all features from visible input are retained. Subsequently, we analyze feature maps from global attention outputs. These outputs contain features from all modalities input. These features func-

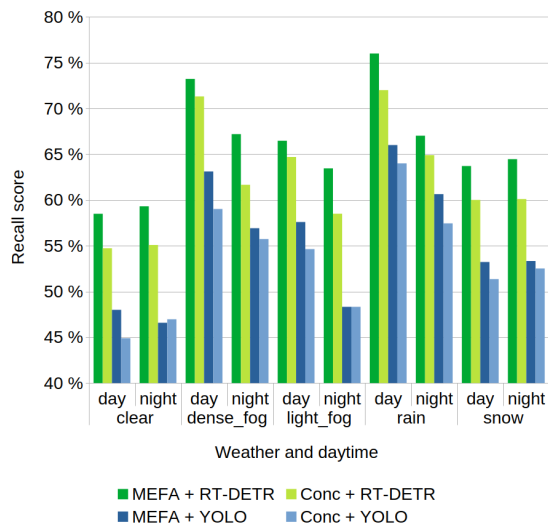


Figure 3: Recall performance according to weather and daytime conditions. Models are RT-DETRv1 combined with MEFA model (MEFA + RT-DETR) and with channel-wise concatenation (Conc + RT-DETR); YOLOv8 model combined with MEFA module (MEFA + YOLO) and with channel-wise concatenation (Conc + YOLO).

tion as a filter to pick up relevant visible features that play a determinant role during feature extraction of object detection models.

5.2.2 Final Visual Results

Figure 5 illustrates the results obtained in two distinct weather conditions. The first two lines present an example of the input in rainy conditions, while the subsequent two lines illustrate an input in clear night conditions. To demonstrate the potential limitations of the channel-wise concatenation approach, the confidence level of objects was set to 0.25, thereby highlighting the occurrence of intriguing false positive pedestrian detection. In the first example, a false positive pedestrian has been identified by the channel-wise concatenation approach in the middle of the image. The second example illustrates how the model, when combined with MEFA, is more effective in detecting vehicles on the right of the image that are discernible only in the LiDAR input image. The intermediate fused image output of the first example is illustrated in Figure 4.

6 DISCUSSION

The MEFA module demonstrates superior accuracy when integrated with a one-stage object detection model, YOLOv8, or a transformer-based model, RT-DETRv1, compared to the channel-wise concatena-

tion early fusion approach with the same model. The module uses an attention mechanism similar to the mid-fusion scheme. This mechanism allows the module to combine relevant features from each modality by using local attention in the initial stage and then filter these fused features with attention computed on each modality feature, which acts as global attention. We hypothesize that an attention mechanism is essential to enable the model to distinguish which features are critical for detection and which features must function as filters. This hypothesis enables the creation of interactions between the multimodal features, thereby improving accuracy performance, particularly in adverse weather conditions. In such conditions, sensors can provide noisy data leading to false detections. It also enables robust detection when sensors are unable to provide information due to time of day or weather conditions.

The output of the MEFA module is a feature map that functions as a single modality image. This property allows the utilization of state-of-the-art single-modality models that are becoming increasingly prevalent in the computer vision field. This aspect also permits more rapid training and enhanced performance using pre-trained weights derived from alternative datasets. Furthermore, by parameterizing the final convolution block to output features as a three-channel image, it is now possible to employ black-box models not able to accept images with more than three channels. This feature has also prompted new investigations into the potential influence of this parameter on object detection outcomes.

Nevertheless, it is essential to acknowledge the inherent constraints of this approach. The module introduces additional memory and computational constraints and operates at a processing speed of 200 milliseconds, in comparison to the object model, which operates at 5 milliseconds for YOLOv8 and 35 milliseconds for RT-DETRv1¹. Further ablation studies could be conducted to gain a deeper understanding of the role of the attention mechanism and whether alternative strategies, such as Transformer-based attention, could improve performance. Additionally, it is crucial to recognize that data transformation, particularly in the case of LiDAR sensors, can result in information loss.

¹The reference speed value was obtained on a Dell Precision 5770 with an Intel® Core™ i7-12800H CPU, a NVIDIA RTX A3000 12 GB GDDR6 GPU, and 32 GB of DDR5 RAM.



Figure 4: Example of intermediate fused image output from MEFA module combined with RT-DETRv1 model. Grayscale images correspond to a specific single channel of the color image.

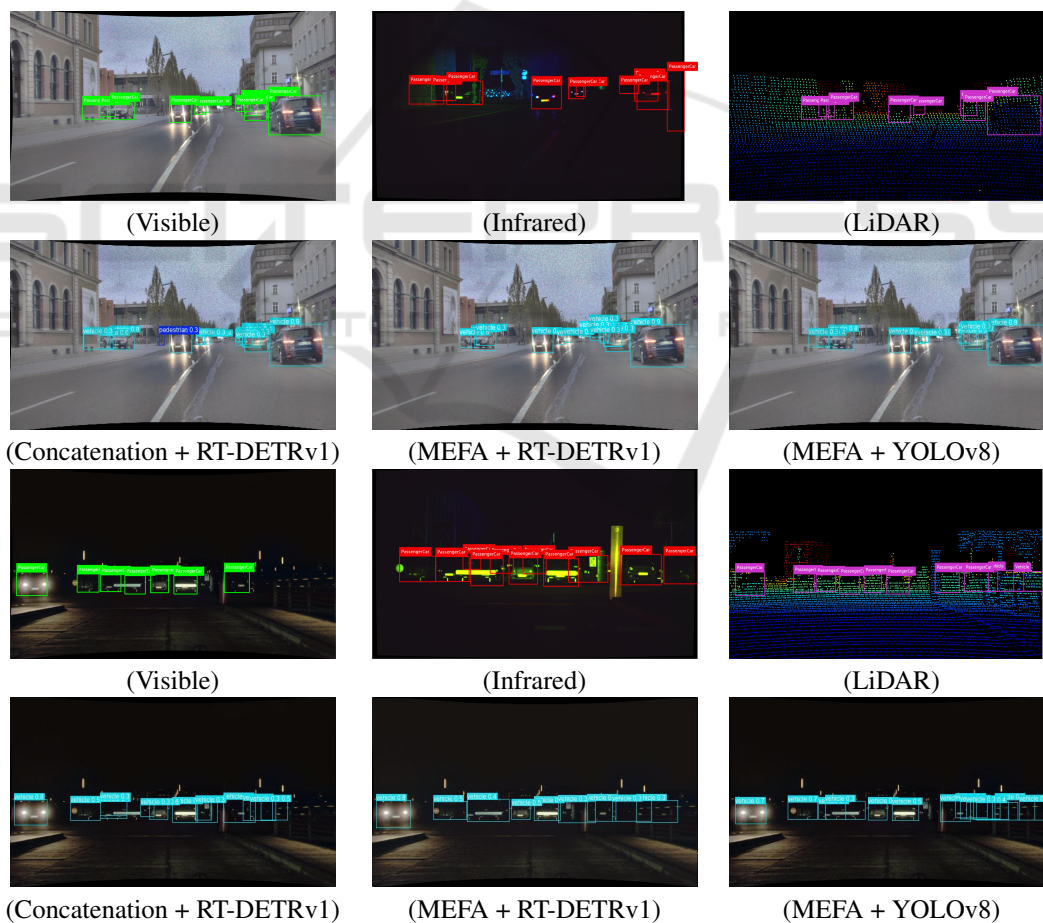


Figure 5: Examples of rain and clear night images. *Visible*, *Infrared* and *LiDAR* indicate visible, infrared and LiDAR sensor images respectively with their ground truth labels. *Concatenation + RT-DETRv1*, *MEFA + RT-DETRv1* and *MEFA + YOLOv8* indicate the channel-wise concatenation or the MEFA module with the RT-DETRv1 model and the MEFA module combined with the YOLOv8 model. Light blue and dark blue indicate vehicle and pedestrian detection respectively.

7 CONCLUSION AND PERSPECTIVES

This article presents a novel early fusion approach based on our MEFA module. The MEFA module, combined with state-of-the-art models, improves, especially in adverse weather conditions, the performance accuracy of vehicle and pedestrian multimodal detection. Furthermore, the MEFA module can improve any single modality model, especially a black box model, for any multimodal application.

In terms of future research, we identified several potential avenues. Firstly, optimizing the module architecture could reduce the computational load, especially when dealing with features of large spatial dimensions. Additional sensor types integration, such as radar or ultrasonic sensors, would be beneficial in investigating and improving detection robustness in challenging conditions. Secondly, further research could be carried out on the MEFA module to better understand the impact of characteristics of modalities and external factors, such as weather or visibility, on the accuracy.

In light of climate change, we aim to direct our future efforts toward enhancing the module to minimize its energy consumption and evaluate the carbon footprint of our models. Furthermore, we intend to investigate the integration of our model into edge devices, exploring innovative approaches to optimize performance while maintaining sustainability. It would be a question of conducting holistic research considering the dimensions of (a) measurements and estimations, (b) algorithms, methods, and models, (c) extreme edge, and (d) understanding the systemic effects of AI.

ACKNOWLEDGEMENTS

This work was carried out in part within the framework of the "Edge Intelligence" Chair within MIAI of the University of Grenoble Alpes, project referenced ANR-19-PIA3-0003.

REFERENCES

- Bijelic, M., Gruber, T., Mannan, F., Kraus, F., Ritter, W., Dietmayer, K., and Heide, F. (2020). Seeing Through Fog Without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen Adverse Weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11682–11692.
- Chaturvedi, S. S., Zhang, L., and Yuan, X. (2022). Pay "attention" to adverse weather: Weather-aware attention-based object detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4573–4579.
- Chen, Y.-T., Shi, J., Ye, Z., Mertz, C., Ramanan, D., and Kong, S. (2022). Multimodal object detection via probabilistic ensembling. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T., editors, *Computer Vision – ECCV 2022*, pages 139–158, Cham. Springer Nature Switzerland.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houtsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.
- Grauer, Y. (2014). Active gated imaging in driver assistance system. *Advanced Optical Technologies*, 3(2):151–160.
- Huang, K., Shi, B., Li, X., Li, X., Huang, S., and Li, Y. (2022). Multi-modal Sensor Fusion for Auto Driving Perception: A Survey.
- Jocher, G., Chaurasia, A., and Qiu, J. (2023). Ultralytics YOLO.
- Malawade, A. V., Mortlock, T., and Faruque, M. A. A. (2022). Ecofusion: Energy-aware adaptive sensor fusion for efficient autonomous vehicle perception.
- Martínez-Díaz, M. and Soriguera, F. (2018). Autonomous vehicles: Theoretical and practical challenges. *Transportation Research Procedia*, 33:275–282.
- Stahlschmidt, S. R., Ulfenborg, B., and Synnergren, J. (2022). Multimodal deep learning for biomedical data fusion: A review. *Briefings in Bioinformatics*, 23(2):bbab569.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Tabassum, N. and El-Sharkawy, M. (2024). Vehicle detection in adverse weather: A multi-head attention approach with multimodal fusion. *Journal of Low Power Electronics and Applications*.
- Terven, J., Córdova-Esparza, D.-M., and Romero-González, J.-A. (2023). A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine Learning and Knowledge Extraction*, 5(4):1680–1716.
- Xiang, C., Feng, C., Xie, X., Shi, B., Lu, H., Lv, Y., Yang, M., and Niu, Z. (2023). Multi-Sensor Fusion and Cooperative Perception for Autonomous Driving: A Review. *IEEE Intelligent Transportation Systems Magazine*, 15(5):36–58.
- Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., and Chen, J. (2024). DETRs Beat YOLOs on Real-time Object Detection.