Gated Multimodal Fusion for Visible-Thermal Pedestrian Detection

Amine Chadli, Etienne Balit, Pierre Arquier
Neovision
10 Rue d'Arménie, 38000 Grenoble, France
etienne.balit@neovision.fr

Guillaume Delubac, Emmanuel Bercier, Xavier Brenière
Lynred
364 route de Valence, Actipole CS 10027, 38113 Veurey-Voroize, France
emmanuel.bercier@lynred.com

September 29, 2020

Abstract

Robust detection of pedestrian in challenging illuminations is essential for autonomous driving. Visible and thermal cameras have been shown to be complementary for detecting pedestrian under day and night settings.

In this paper, we present an internal multimodal dataset for pedestrian detection. This dataset is composed of image pairs taken in the visible and the far infrared spectrum. The infrared images showcase the image quality provided by Lynred's state-of-the art sensors.

Moreover, we design a new approach to visible and thermal fusion for pedestrian detection based on state-of-the-art methods for multimodal fusion and object detection. Our method uses multi-task learning to train an end-to-end model to perform detection on the multimodal information and on each modality simultaneously, we also introduce a gating mechanism to actively weight each modality per pixel.

We evaluate the improvement brought by each component of our approach on the introduced multimodal dataset for pedestrian detection, and we compare its performances to multiple baselines. Our fusion method outperforms other fusion techniques and it achieves the best performance compared to early and late fusion methods. Our results show that the addition of infrared sensors to the cameras operating in the visible spectrum leads to a significant improvement in pedestrian detection thanks to the thermal signature detection and that all sensors can be easily implemented into standard ADAS platforms.

Keywords— Autonomous Driving, Far-infrared, Pedestrian Detection, Gated Multimodal Fusion, Multitask Learning, Deep Neural Networks

1 Introduction

More than 1.3M people are killed every year due to road fatalities. As presented in table 1, the French Road Safety Observatory [9] details figures of fatalities by category of road users. If car user and motorcyclist fatalities represent the largest part of road fatalities, pedestrian and pedal cyclist fatalities represent 21%, and have increased over the last six years. The French Road Safety Observatory also reports that 70% of the seriously injured people are vulnerable road users (VRU) such as motorcyclists, moped users, cyclists and pedestrians.

In its 2025 roadmap, the Euro NCAP organization [8] reports that VRU injuries reach more than 135 000 people in Europe each year. ADAS are expected to have a key impact on reducing VRU injuries and fatalities. In

	Fatalities 2016	Variation 2010-2016
Car users	51%	-17%
Motorcyclist	18%	-13%
Pedestrians	16%	+15%
Pedal cyclists	5%	+10%

Table 1: 2016 French accidentology data

the set of ADAS functionalities, the SDB[1] underlines the impact of the Autonomous Emergency Braking (AEB) that could decrease VRU fatalities by 13% to 18% in the United States.

AEB is widely deployed into Euro NCAP roadmap [8] in AEB VRU, cyclist, pedestrian, junction & crossing and Head-on. Before planning decision and performing autonomous braking, AEB should rely first on VRU and/or obstacle robust detection.

In a study leaded by the AAA in October 2019 [1], four vehicles were selected (midsize sedans from 4 different manufacturers) and tested according to euro NCAP scenario. The four selected vehicles had pedestrian detection system with collision mitigation functionality. Research questions were:

1. How do vehicles equipped with pedestrian detection systems perform when encountering an adult pedestrian crossing the roadway in different situations?
   Results: Even in the simplest situation at 20mph, reliability of systems is not perfect as collision with an adult pedestrian target was avoided 40% of the time, and during an additional 35% of the time collision were mitigated by an average speed of 4.4mph. For the last 25% of time, vehicles impacted the pedestrian target at full speed.

2. How do pedestrian detection systems function at night?
   Results: Evaluated pedestrian detection systems were ineffective under nighttime conditions.

These results show that improvements are still required for the AEB function in normal conditions and that new solutions are needed in order to address challenging scenarios, especially nighttime conditions where none of the four vehicles has detected any pedestrian.

In order to sense and to understand the surroundings of the vehicle, sensors based on four different technologies are mainly considered: Ultrasonic, RADAR, LIDAR and Cameras. Because all of these technologies have advantages and drawbacks regarding detection performance, a sensor fusion will enable to optimize perception performance in line with recommendations of Dibotics [3]. Regarding imaging technology used in camera sensors, the AWARE project [10] went to the conclusion that Visible RGB extended to NIR (or Red-Clear sensors) combined with LWIR (or Far-Infrared or Thermal) provide the best spectral bands combination to improve ADAS detection performance of vehicle, pedestrian, bicycle, animals or road marking, and recognize traffic signs, in all weather conditions.

This paper focuses on far-infrared camera integration in combination with a RGB camera, presents an internal color-thermal dataset, introduces a deep learning strategy to improve pedestrian detection, and presents our results obtained with different fusion approaches.

2 Thermal camera detection principle

According to the Plank law presented in the Figure 1, any object will radiate light energy mainly depending on its own temperature and its properties of emissivity.

Simplified equation of the Planck is the Wien displacement law (Figure 2).

With λ being the spectral wavelength and T the temperature of the object. Wien law gives the maximal wavelength at which the object will radiate according to its temperature. Thus human or any VRU with a temperature closed to $300K$ will radiate their energy around a wavelength pic around 10 μm. Figure 3 details the different spectral bandwidth with respect of the infrared light, from Short Wave to Very Long Wave including Mid and Long wave.

Microbolometers MEMs' based sensor convert long-wave infrared radiation into current. To be sensitive only to the light flux coming from the scene, microbolometer

[1] SDB: https://www.sbdautomotive.com

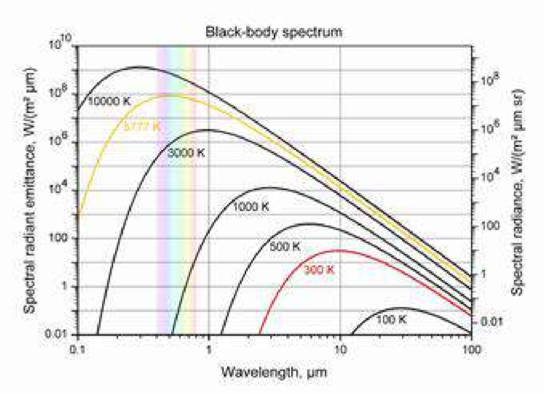


Figure 1: Planck law.

$$\lambda_{\max}(T) = \frac{2.89777291 \cdot 10^{-3} \cdot K}{T}$$

Figure 2: Wien law.

sensors are packed into vacuum package in order to isolate pixels from the surroundings. Performance of microbolometer based sensor will mainly depend on vacuum package quality, pixel design, thermometer material and CMOS readout integrated circuit used to bias microbolometer and arrange pixel data. Over a large operating temperature from -40°C up to +85°C, state of the art thermal sensors based on 12μm pixel pitch microbolometer can typically detect 50 mK of temperature variation and perform image size from 80x80 to 1280x1024 pixels at frame up to 120 frames.

Thus, microbolometer generates thermal images only depending on object temperature and structure, constant and immune to any external light source such as sun, head or street lights.

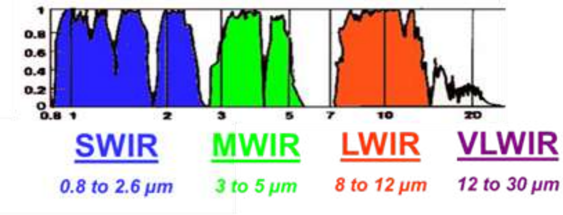


Figure 3: Imaging spectral bandwidth.

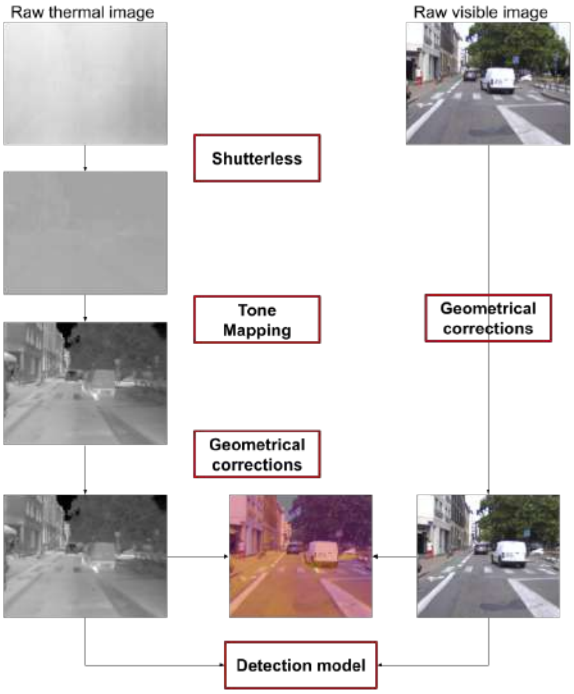


Figure 4: From raw to processed images

3 Cameras integration

3.1 Cameras setting

The full system consists of two pairs made of an infrared camera and an RGB camera, mounted on an aluminium structure.

The infrared cameras integrate a VGA sensor (640x480 pixels) with a lens allowing an horizontal field of view of 42° and an aperture of $f/1.2$.

The RGB cameras integrate a Sony IMX273 sensor (1448x1086 pixels) with a lens allowing an horizontal field of view of 45° and an aperture of $f/1.4$.

The cameras on a pair are 12.7 cm apart and the two pairs are 34.1 cm apart. The main RGB camera controls the trigger, that is sent to the three other ones, to synchronize the capture at 30 frames per second. The

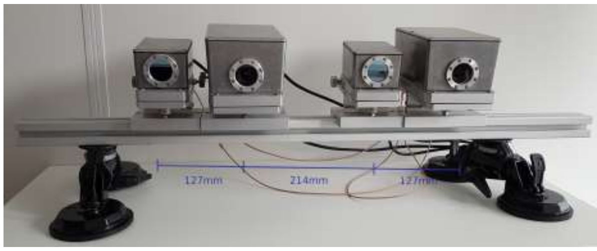


Figure 5: Camera system

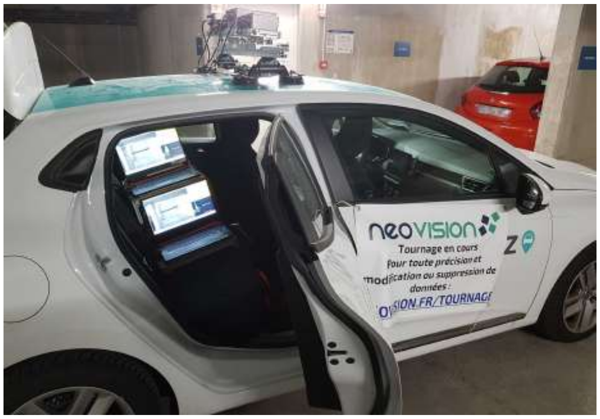


Figure 6: Full system mounted on the car

system is installed on top of a car. The cameras where initially adjusted to look at a common point roughly 5 meters in front and 1 m below the structure.

Each pair of cameras is sending its images to a dedicated laptop, running a piece of software that records them in separated image files, with lossless compression.

Infrared images are recorded as 640x480 16bits images. Images from the RGB camera are initially 1448x1086 image but only the 1280x960 rightmost and vertically centered region is kept (cropped in the sensor), to facilitate alignment, as it corresponds roughly to the region seen by the infrared camera.

3.2 Geometric calibration

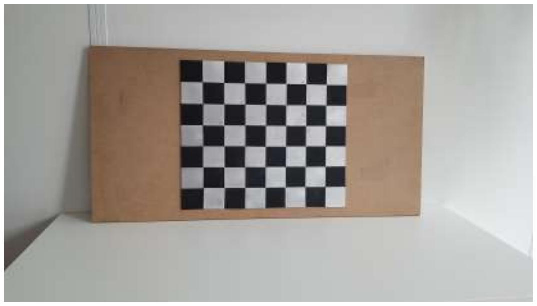


Figure 7: Chessboard used for calibration.

Camera calibration usually involves the use of a chessboard. However a simple chessboard isn't visible by thermal cameras. In an effort to calibrate at the same time infrared cameras and RGB cameras as well as to facilitate cameras alignment, we crafted a chessboard visible in both spectra.

It consists in a wood plate on which 5 cm wide squares are glued forming a 7x8 chessboard pattern. Black squares are cut in a 5mm thick foam board and white squares are cut in 5 mm thick aluminium ruler. Aluminum squares are sanded to avoid specular reflection.

When used with a correct angle, the corners of the chessboard can be detected in RGB images using OpenCV directly. Detection is possible in infrared images thanks to the reflection of a uniform body's radiation by the aluminium, giving a different reading for the black squares and white squares temperatures. Usually the body used is the roof, constraining the angle of the chessboard used.

However, corners detection in infrared images is not as easy as in RGB. Images are corrected subtracting a shutter image. Every images intended to be used in calibration had the approximate four outer corners of the pattern labeled manually to entirely avoid the chessboard detection step. Corners detection is made by warping the labeled chessboard into a rectangle, and filtering of the image before using OpenCV corner detection, due to the heavy noise of the infrared image.

The filtering step consists in computing the median value μ in the warped area, clipping the image between the median of the values below μ and the median of the values above μ. This step removes a lot of the noise, and separates the white squares from the black nicely. A Gaussian blur and a median blur with 3x3 kernels are finally applied. This gave the best results with OpenCV detection.

The pandemic context made it very difficult to find the material and tools to craft the board, hence a result not very accurate, but it was sufficient for our application.

3.3 Thermal camera signal enhancement

Far-infrared cameras are sensitive to any temperature change either coming from the scene of interest or coming from camera operating temperature drift. The image model of such device is :

$$Output_{i,j} = G_{i,j}(T_{FPA}) \ . \ Scene_{i,j} + O_{i,j}(T_{FPA})$$

Where G and O are respectively a gain and an offset image depending on the sensor temperature T_{FPA}, and inducing non uniform noise on the image. Non uniformity correction (NUC) must be applied.

The most widely used NUC method is the use of a mechanical shutter to observe several frame of a uniform body and deduce the current gain and offset images. However this process must be performed frequently, as the temperature evolves quickly, interrupting the video flow. Moreover the need of a mechanical shutter is major drawback.

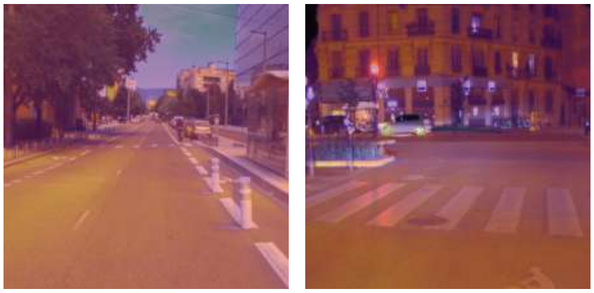
Our choice for NUC is the use of a scene-based shutterless correction algorithm. This method requires a factory calibration using blackbodies of uniform temperature, but doesn't require any more images taken in a controlled environment afterwards.

Additionally we apply a tone mapping to be able to display pixels (16b format) on a screen (8b format). This tone mapping enhance the contrast of the output images while preserving the dynamic.

3.4 Thermal camera alignment with RGB camera

Obtaining both far-infrared and visible image of the same scene presents many interests:

- Domain comparison for challenging tasks such as pedestrian detection.
- Acquisition of multispectral dataset.
- Transfer of meta-data such as labels in machine learning oriented tasks.



(a) Day (b) Night

Figure 8: Examples of RGB-Thermal composite image in day and night conditions.

Digital shift was applied to both images in order to get a perfect alignment of the center of each image. This technique proves its effectiveness for targets far enough compared to the distance between the two cameras. As the two cameras have slightly different fields of views, distortions are different and need to be corrected. The last operation was cropping to the biggest region covered by both images. As a result infrared images were cropped from 640x480 to 555x479 , RGB images were cropped from 1448x1086 to 1117x962. Figure 4 presents the pre-processing steps from the raw images to the input of the detection model. Figure 8 show some examples of aligned images.

4 Dataset description

4.1 Data annotation

8820 frames from the captured videos were sampled for manual annotation, with a maximum of two sampled frames per second of video. We filtered out the images after annotation to only keep the images that have at least one bounding box (5720 images). The training and test videos were selected from different shooting sessions to ensure that our model is tested on images not seen during training.

To annotate the ground truth, we used the CVAT computer vision annotation tool. The annotations were done on the visible frames, while using the paired infrared frames for verification. This was necessary for night time annotation where distant pedestrians are rarely visible in the color images.

Individual pedestrians and people riding a two-wheeled vehicle were labeled as ‘person’. Not distinguishable individuals were labeled as ‘crowd’. In the training set, we add the ‘occlusion’ tag to specify occluded pedestrians. In the testing set, we separate two types of occlusion : partial occlusion (less than 50% of the object occluded) and heavy occlusion (more than 50% of the object occluded). Moreover, objects that could not be clearly identified as persons by annotators were labeled as ‘person?’ and are ignored in the evaluation.

4.2 Dataset properties

Different shooting sessions were done different day and night times. The overall makes a total video duration of about 9 hours. Table 2 present the number of acquired frames corresponding to each shooting scenarios, and the total time of the scenario in the overall videos.

To show the diversity of our data, we also present the distribution of the annotated data according to the illumination (day/night) setting, occlusion type, and the distance of objects from the camera.

- **Dataset distribution according to the day/night setting:** We made sure to have a balanced distribution between day and night. Table 3 presents the number of images and the number of bounding boxes for each setting.

- **Dataset distribution according to occlusion type:** The histogram in figure 9 shows the distribution of the test dataset according to the amount of occlusion.

- **Dataset distribution according to distance from camera:** Figure 10 shows the distribution of the test dataset (only not occluded objects) according to the distance from the cameras. We mention that the annotation were very delicate, hence very small (far) objects were also annotated, we set 19 pixels as the lower limit of the bounding box diagonal size.

Settings	Images number	Appearance time (s)
Day clear extra urban	44971	1499
Day clear parking lot	47451	1581
Day clear school	1947	141
Day clear tunnel	3824	136
Day clear urban	434698	16168
Night clear extra urban	6464	215
Day clear urban	369159	12305
Total classified	908514	32045

Table 2: Shooting scenarios of the overall acquired data

	Training		Testing		All
	Day	Night	Day	Night	
# Images	2126	2083	730	781	5720
# Annotations	8293	8090	3132	2480	21995

Table 3: Dataset distribution according to day/night setting

5 Pedestrian detection

The task of pedestrian detection is well studied by the computer vision community. This is a challenging problem because of the variety of poses and appearances that a pedestrian can take and the large number of possible use cases. Most approaches in the literature use Convolutional Neural Networks (CNN) for pedestrian detection, applied to visible images from RGB cameras. However, using only the visible information has some limitations such as insufficient lighting sensitivity. Thermal cameras are more robust to these conditions since they are sensitive to far infrared radiations directly emitted by objects.

In a previous work [2], we showed that the combination of infrared and thermal cameras could improve the performance of a state-of-the-art pedestrian detection method even when using a basic method for fusing both modalities. In this work, we are especially interested in studying how to fuse both modalities to maximize the robustness of a pedestrian detection system.

Recent work have proposed new Deep Neural Network architectures for multimodal fusion to learn the best way to fuse the modalities based on the data. CentralNet [11] proposed a meta-architecture in which a

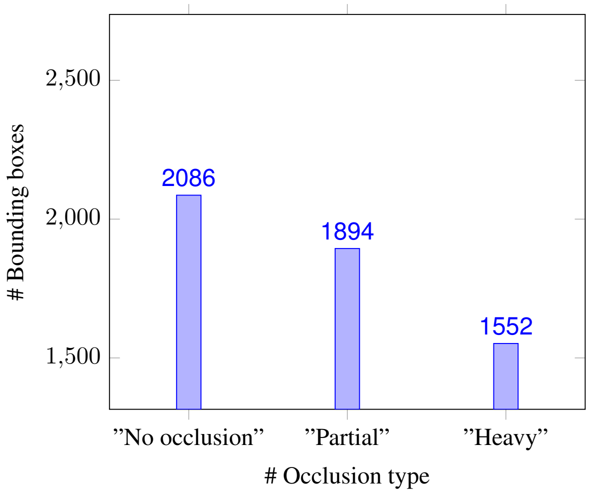


Figure 9: Test dataset distribution according to occlusion type.

central pathway draws input as a weighted sum of each modality at multiple level of the two unimodal Deep Neural Network. The weights for this multimodal fusion are parameters of the models that are trained end-to-end with the rest of the architecture. In addition, the authors use a multi-task learning scheme to maximize the performance of the central pathway while keeping good performances for the unimodal networks. Other work propose to adaptively modulate the fusion based on the input image. In the domain of object detection, GIF networks [6] proposed a gating mechanism composed of a weight generator module that takes the concatenation of the features from both modalities as input and predict a weight for each pixel and each modality.

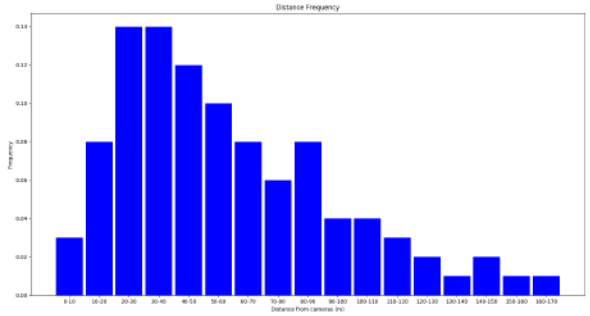


Figure 10: Test dataset distribution according to distance from the camera.

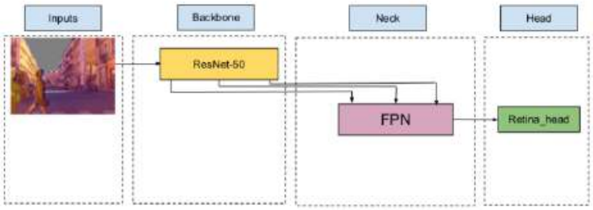


Figure 11: Early fusion with composite images

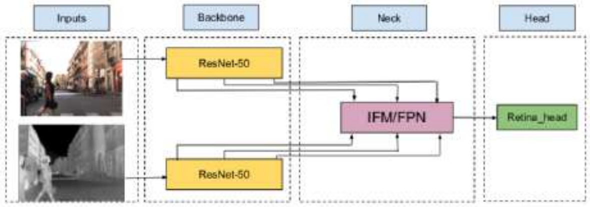


Figure 12: Intermediate Gated Fusion

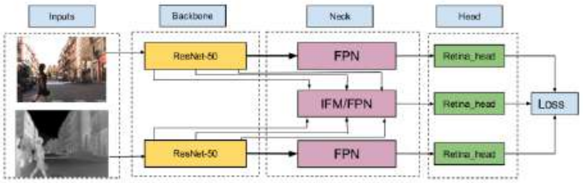


Figure 13: Multi-Task Training

5.1 Methodology

We integrate ideas from both CentralNet and GIF networks to design an object detection architecture that can learn from data how to fuse features from visible and far infrared images. Building on our previous work, we use RetinaNet [7] as a basis for this work. RetinaNet is part of a family of architectures called "single shot detectors" (SSD). The principle of this type of architecture is to predict if there is an object of a given class in a large number of fixed bounding boxes called "anchor boxes". Those boxes of multiple scale and aspect ratio are distributed on a grid on the image. The network is also trained to predict a correction of those boxes to better fit the object. SSD are composed of a pretrained backbone that extract features from the input images, a neck that integrate features of multiple levels together and a head that output the prediction for each anchor box.

Our proposed architecture is composed at training time of two auxiliary RetinaNet, one for each modality, and a central pathway for the multimodal fusion composed of an head and a modified neck. This additional pathway is the only one required during inference. We modify the Feature Pyramid Network (FPN) neck module from RetinaNet to integrate an Information Fusion Modules (IFM) for each level of the pyramid. The IFM takes as input the concatenation of features from both modalities and generate a per-pixel weight map for each modality. It then feed the dynamically weighted sum of the features from both modalities to the FPN.

5.2 Training setup

In order to fairly compare the different domains under the same conditions, the architecture is kept the same (RetinaNet with FPN) while varying the fusion method. Hyper-parameters and data augmentation scheme are also kept the same for all experiments. In the intermediate multimodal fusion experiment, a specific multimodal augmentation strategy is added. As the architecture takes RGB images as input, far-infrared images were converted to RGB images using the inferno colormap [5], a perceptually-uniform colormap that better use the three color layers. Images were also resized to match the visible images size.

The RetinaNet model is initialized using weights pretrained on MS COCO dataset and trained with Stochastic Gradient Descent optimizer (SGD) with a batch size of 2, a momentum of 0.9 and a weight decay of 0.0001. A step decay scheme is used to gradually decrease the learning rate with an initial learning rate of 0.01.

5.3 Experiments

- Trained on visible only: RetinaNet model trained on the RGB images.

- Trained on infrared only: RetinaNet model trained on the thermal images.

- Early fusion: RetinaNet model trained on the composite images.

- Late fusion: Models trained on each modality are used to predict raw detections on each modality respectively. These detections are then combined before applying the 'NMS' algorithm to produce the final detections.

- Intermediate Gated fusion: Intermediate fusion consists of having a single model for both modalities. In this case, two backbones are responsible for extracting the feature maps of the corresponding modality. These features are then passed through a multimodal fusion module which merges the features of both modalities at each level of the backbone, before transmitting them to the FPN. Finally, the set of features after fusion is used to predict detections.

5.4 Evaluation and results

To evaluate the results, we use both Average Precision (AP) and Log Average Miss Rate metrics. Average Precision (AP) corresponds to the area under the Precision-Recall curve. An AP of 1 corresponds to perfect detection on all the dataset, as computed against the human annotated ground-truth. A "detection" is considered as correct when the predicted bounding box is similar to the annotated one. Intersection over Union (IoU) is used as a similarity metric, corresponding to the ratio of the intersection of the two bounding boxes over the union of those. The evaluation presented in Tables 4 and 5 is done with an IoU threshold equal to 0.5. The Log-Average Miss Rate (LAMR) referenced in [4] is calculated by averaging miss rate at nine False Positive Per Image FFPI rates evenly spaced in log-space in the range 10^{-2} and 10^{0}.

As shown in these results, far-infrared and visible excel in different conditions: the visible modality performs the best during the day while far-infrared performs the best during the night.

Contrast between pedestrians and their surrounding is the key element for detection. For visible images, maximum contrast corresponds to a situation with high luminosity also being the warmest of the day. At the opposite

Methods	Illumination			Occlusion			Distance			
	Day	Night	Both	None	Partial	Heavy	Near	Medium	Far	Very far
							<10m	<30m	<50m	>50m
Infrared only	36	62	47.2	61.5	35.4	13.7	93.9	91.2	75.4	24.3
Visible only	**67**	48.5	58.9	71.4	49.1	21.8	**99.9**	97	90.2	38.3
Early fusion	58.8	57.3	58.1	71.8	47.6	19.7	98.6	97	90.5	38.3
Late fusion	62.2	62	62	74.5	51.9	22.7	99.6	97.8	93.5	42
Gated fusion	66.3	**66.1**	**66.2**	**78.1**	**58.5**	**28.3**	99.6	**97.9**	**94.7**	**50.4**

Table 4: AP results on the test dataset of different methods in different settings

Methods	Illumination			Occlusion			Distance			
	Day	Night	Both	None	Partial	Heavy	Near	Medium	Far	Very far
							<10m	<30m	<50m	>50m
Infrared only	78.8	63.7	72.3	55.9	77.6	92.7	11.1	17.1	40.6	86.2
Visible only	60	70.2	64.3	55.9	77.6	92.7	0	5.8	19.6	78
Early fusion	65.5	65.6	65.4	44.7	69.9	89.3	2.3	6.5	19.8	78.7
Late fusion	65.1	63.5	65.4	45.7	70.7	90.5	0	4.5	15.1	77.6
Gated fusion	**57.5**	**58.3**	**57.9**	**38.2**	**62.3**	**85.7**	**0**	**3.8**	**11.1**	**70.4**

Table 5: LAMR results on the test dataset of different methods in different settings

for far-infrared images, maximum contrast corresponds to lower temperature of the road and building, like night or bad weather conditions. Far-infrared and visible images are thus complementary for pedestrian detection.

The complementary nature of visible and far-infrared is reinforced by the good performances of the gated multimodal fusion network as it obtains the best overall results. Some examples of complementary detections on both day and night in regards to different difficult situations (far distance, poor illumination, and liveness difficulty detection) are shown in the Appendix.

6 Conclusion

In this paper we introduced a novel internal visible-thermal dataset for pedestrian detection. We used two visible-thermal camera pairs for the dataset shooting. We expect to exploit the stereo information in future work.

We've given details about the data distribution according to different aspects such as : illumination, occlusion type and distance from the camera. For the illumination aspect, our dataset is pretty balanced between day and night. As for the occlusion type aspect, we specify 3 occlusion tags (no occlusion, partial occlusion and heavy occlusion), this enables an advanced analysis of the pedestrian distribution, but it's also necessary for the detection results interpretation. As regards to the distance from the camera, the annotated pedestrians are distributed in a large interval (0-170m), in fact very far pedestrians were annotated to be able to analyse the behaviour of the detection model at long range.

Moreover, we presented a new multimodal fusion method for pedestrian detection. This method uses multitask learning and a specific data augmentation strategy to train a robust end-to-end model for pedestrian detection. In a previous work [2], we studied the use of an early fusion method that improved the overall detection results but gave worse results than the ‘only_visible’ method during daytime and than ‘only_thermal’ during nighttime.

With the use of the gated fusion,

- in day condition, average precision become competitive with only_visible method while LAMR is improved.
- in night condition, both metrics are improved.
- the gated fusion approach outperforms all other tested fusion approaches.
- It has a large gain compared to the use of only the visible information, especially in adverse conditions (night, headlight glare, occlusions, long range, entrance/exit of tunnel).

References

[1] AAA. Authomatic Emergency Braking With Pedestrian Detection. 2019.

[2] E. Bercier, B. Louvat, O. Harant, E. Balit, J. Bouvattier, and L. Nacsa. Far-infrared thermal camera: an effortless solution for improving ADAS detection robustness. In Michael C. Dudzik and Jennifer C. Ricklin, editors, *Autonomous Systems: Sensors, Processing, and Security for Vehicles and Infrastructure 2019*, volume 11009, pages 21 – 30. International Society for Optics and Photonics, SPIE, 2019.

[3] Bravo. Using LIDAR asd the Amygdala of the Self-Driving car. 2018.

[4] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012.

[5] S Garnier. viridis: Matplotlib default color map. *R package version 0.2*, 3, 2015.

[6] Jaekyum Kim, Junho Koh, Yecheol Kim, Jaehyung Choi, Youngbae Hwang, and Jun Won Choi. Robust Deep Multi-modal Learning Based on Gated Information Fusion Network. *arXiv e-prints*, page arXiv:1807.06233, July 2018.

[7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *arXiv e-prints*, page arXiv:1708.02002, August 2017.

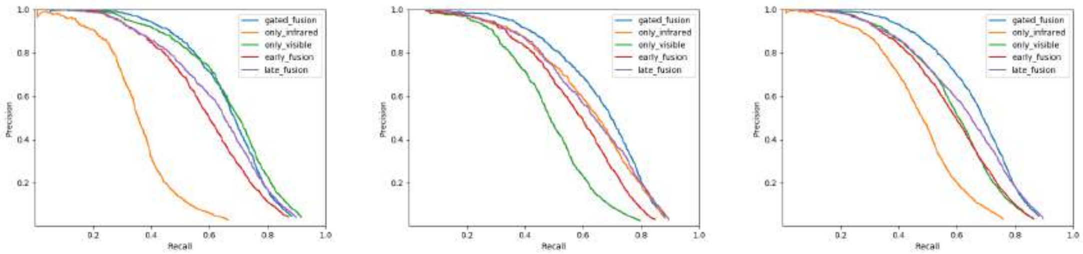
[8] Euro NCAP. Roadmap 2025-V4. 2018.

[9] French Road Safety Observatory (ONISR). Road safety in 2016. 2016.

[10] Nicolas Pinchon, Olivier Cassignol, Adrien Nicolas, Frédéric Bernardin, Patrick Leduc, Jean-Philippe Tarel, Roland Brémond, Emmanuel Bercier, and Johann Brunet. All-weather vision for automotive safety: which spectral band? In *International Forum on Advanced Microsystems for Automotive Applications*, pages 3–15. Springer, 2018.

[11] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

7 Appendix

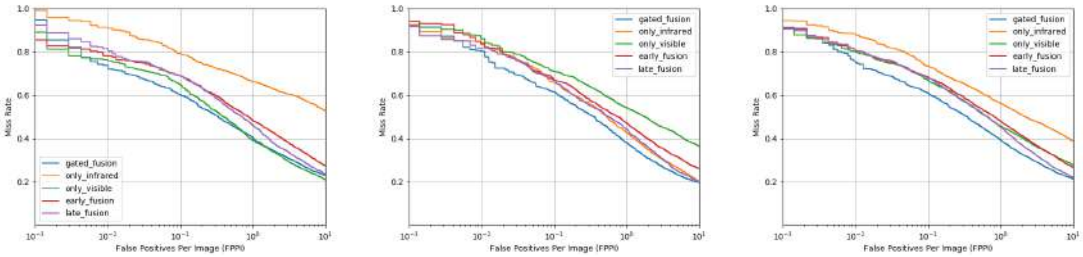


(a) day

(b) night

(c) both

Figure 14: Precision-Recall curves.



(a) day

(b) night

(c) both

Figure 15: Miss Rate vs False Positive Per Image curves.

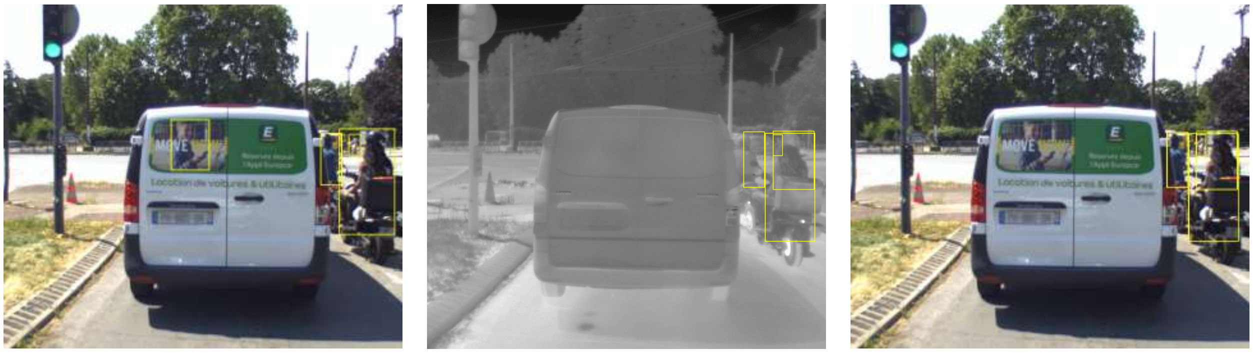


Figure 16: Difficult day example due to liveness detection difficulty. Left: results using only the visible modality. Right: results using our gated fusion method.

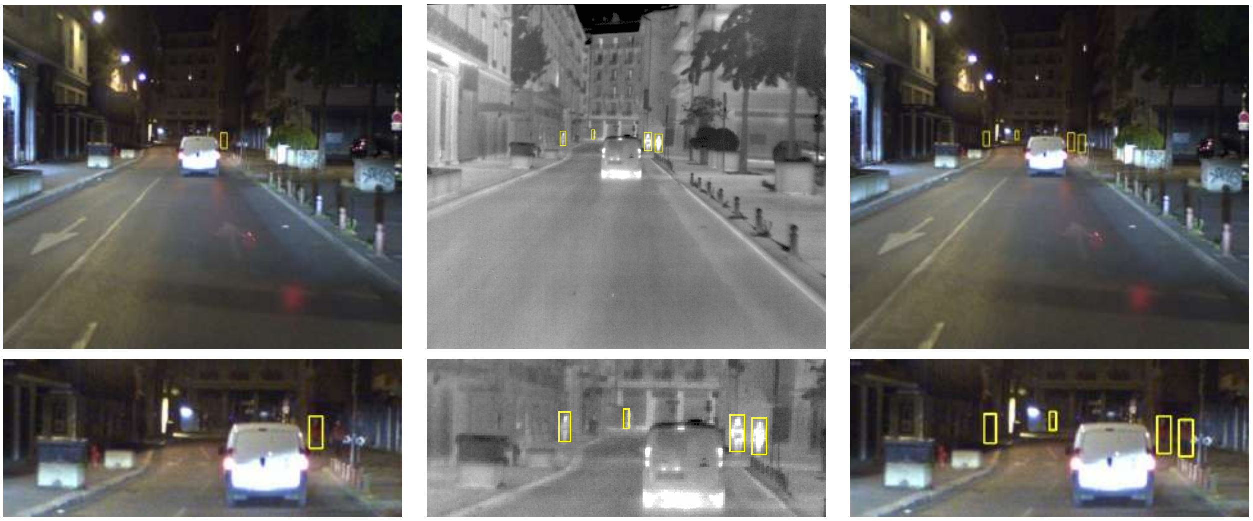


Figure 17: Difficult night example due to long distance from camera. Top row: full image. Bottom row: zoomed on the region of interest. Left: results using only the visible modality. Right: results using our gated fusion method.

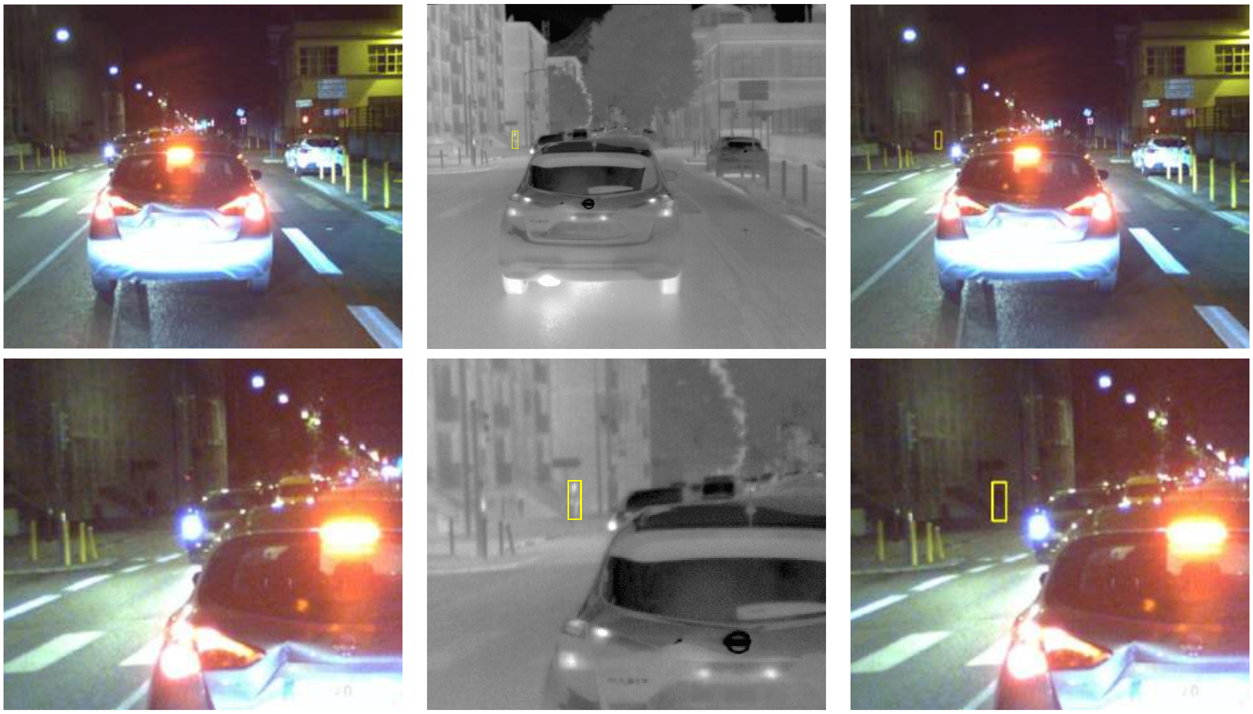


Figure 18: Difficult night example due to poor illumination. Top row: full image. Bottom row: zoomed on the region of interest. Left: results using only the visible modality. Right: results using our gated fusion method.