sensors

Article

End-to-End Detection of a Landing Platform for Offshore UAVs Based on a Multimodal Early Fusion Approach

Francisco Soares Neves [1,2], Rafael Marques Claro [1,2] and Andry Maykol Pinto [1,2,*]

[1] Faculty of Engineering, University of Porto (FEUP), 4200-465 Porto, Portugal
[2] Centre for Robotics and Autonomous Systems—INESC TEC, 4200-465 Porto, Portugal
[*] Correspondence: andry.m.pinto@inesctec.pt

Abstract: A perception module is a vital component of a modern robotic system. Vision, radar, thermal, and LiDAR are the most common choices of sensors for environmental awareness. Relying on singular sources of information is prone to be affected by specific environmental conditions (e.g., visual cameras are affected by glary or dark environments). Thus, relying on different sensors is an essential step to introduce robustness against various environmental conditions. Hence, a perception system with sensor fusion capabilities produces the desired redundant and reliable awareness critical for real-world systems. This paper proposes a novel early fusion module that is reliable against individual cases of sensor failure when detecting an offshore maritime platform for UAV landing. The model explores the early fusion of a still unexplored combination of visual, infrared, and LiDAR modalities. The contribution is described by suggesting a simple methodology that intends to facilitate the training and inference of a lightweight state-of-the-art object detector. The early fusion based detector achieves solid detection recalls up to 99% for all cases of sensor failure and extreme weather conditions such as glary, dark, and foggy scenarios in fair real-time inference duration below 6 ms.

Keywords: object detection; sensor fusion; early-fusion; computer vision; RGB camera; thermal camera; 3D LiDAR

check for updates

Citation: Neves, F.S.; Claro, R.M.; Pinto, A.M. End-to-End Detection of a Landing Platform for Offshore UAVs Based on a Multimodal Early Fusion Approach. *Sensors* **2023**, *23*, 2434. https://doi.org/10.3390/s23052434

Academic Editor: Natividad Duro Carralero

Received: 28 January 2023
Revised: 19 February 2023
Accepted: 20 February 2023
Published: 22 February 2023

1. Introduction

Object detection is a perception task that provides the autonomous systems the necessary awareness of the surrounding environment. By extracting features from raw sensory information, it produces meaningful high-level knowledge about surrounding objects (e.g., other vehicles, pedestrians, goal marks, road lanes). In the end, object detection provides semantic representations about the surrounding environment [1]. For detection, the most common on-board sensors used by autonomous systems are two-dimensional/three-dimensional LiDAR, visual cameras, radar, and thermographic cameras. Each sensor has its own advantages and disadvantages that are directly affected by the environmental conditions. Visual cameras alone are the traditional sources of information, however, in recent years, there is a concern for diversifying the set of sensors to increase the robustness of the system for the case of sensor failure against different environmental conditions [2–4]. For example, thermal cameras discriminate really well different objects emitting different temperature intensities. However, on extreme heated environments (e.g., sunny days, hot rooms), the background and foreground becomes indistinguishable and homogeneous. Visual cameras provide colour and texture information, but for foggy, glary, or dark environments, the signal becomes corrupted and noisy. LiDAR sensors extract depth information by analyzing light energy reflected from objects in the surroundings. Nevertheless, in rainy and foggy environments, the reflected information becomes worthless. Hence, for an autonomous system to be reliable, it is a requirement to endow it with a diverse and distinct set of sensor equipment.

Adopting sensor fusion is one of the building blocks to build reliable perception systems. A system of this kind is still able to meaningfully perceive the environment, not only when there are individual failure, but also when environmental conditions corrupt any sensory information making it not reliable. At greater length, sensor fusion is described by combining the information produced by various sensors representing different modalities (e.g., visual, thermographic, RADAR, LiDAR) into a joint representation to produce a less uncertain input. Sensor fusion can happen in three different stages of the detection pipeline [5]: (i) early fusion refers to the combination of multiple input sources into a unique feature vector before feature extraction; (ii) intermediate fusion refers to the process of extracting correlation between features from the joint combination of sensory information inside the perception module and before classification or regression; and (iii) late fusion happens after the classification/regression procedure.

Extracting and combining relevant features from multiple modalities is a challenging process that requires a lot of domain knowledge and it strongly depends on the requirements of the task to solve. Effectively, deep learning technology is already capable of tackling this process [6–8] autonomously. Apart from this deep approaches, when significant correlation across modalities is smoothly inferred, early fusion should be considered [4]. If it is applied, early fusion produces early joint representations directly from raw data and anticipates correlation and redundancy. Several studies have been conducted to find several pair combinations between point cloud, visual, and thermographic into single concatenated representations [9–12]. However, there is a lack of research on combining the three.

In this paper, we contribute by presenting a novel multimodal early-fusion-based perception system that combines visual, thermal, and three-dimensional LiDAR data information to produce reliable detection capabilities against demanding operating conditions such as extreme weather or modality failure. According to the authors' knowledge, there is no equivalent early-fusion method combining point clouds, thermographic, and visual information. This combination of sensors is implemented by a multimodal fiducial ArUco marker called ArTuga (Patent pending (Portuguese Patent Request (PPP) nr. 118328, and European Patent Request (EP) nr. 22212945.4)) proposed by Claro et al. [13]. It enables multimodal detection against several weather conditions for robotic solutions endowing heterogeneous perception systems comprised by visual, thermographic, and LiDAR-based devices. Therefore, it is not possible to replace the marker with an ordinary object given the constructive characteristics of the ArTuga that provide a spatial alignment of certain elements that facilitate the precise and robust detection of the marker. The contributions of this article include:

- Real-time multimodal marker detection that is deployable onboard a UAV;
- Resilient and high-accuracy detection based on early-fusion approaches against sensor failure integrated in the YOLOv7 framework;
- Robustness against demanding weather and operating conditions for extensive experiments using real UAVs landing on a floating platform;
- A new multimodal dataset collected in real offshore and onshore environments during a UAV landing operation, comprised of diverse joint representations of visual, infrared and point-cloud images.

The paper is organized as follows: In Section 2, there is a review of the literature about object detection and sensor fusion. In Section 3, a novel early-fusion methodology is proposed. Section 4 describes the experimental setup. Section 5 exposes the achieved results. Lastly, Section 6 provides the conclusions.

2. Related Work

Object detection (OD) is a computer vision task described by recognizing, identifying, and locating objects of interest within a picture with a certain degree of confidence. OD has gained popularity since it started to fuse with the strong capabilities of neural network technology. OD is mainly subdivided between single-stage and two-stage detection. These two categories compete by finding the optimal trade-off between detection speed and accuracy [14]. Real-time systems require fast and accurate predictions, therefore building real-time object detectors demands solving the former mentioned trade-off.

Two-stage detectors, such as the RCNN [15], fast-RCNN [16], and faster-RCNN [17], follow region proposal principles that focus on the localization of regions of interest in the image (e.g., where objects of interest are located) before performing detection. Firstly, these regions are estimated. Lastly, from these regions, detection is performed. Depending on the detected objects, correspondent confidence probabilities are extracted. These methods focus on achieving high detection accuracy, however, at the cost of being slow. Hence, two-stage detection is not viable for real-time demands. On the other hand, single-stage detectors, such as the you only look once (YOLO) [18] and the single-shot multibox detector (SSD) [19] perform detection and classification in a singular common step. Single-stage detection prioritizes inference speed over accuracy. Naturally, it is suitable to address real-time constraints. Despite inference speed prioritization, these methods achieve at least similar accuracy to two-staged detectors [14]. Within the available single-stage detectors, YOLO is a suitable choice: (i) it is more popular, accessible, and has a stronger documentation than any other method in the literature; (ii) it achieves highly fair detection speeds (up to 45 FPS), and thus it has at least similar accuracy when compared with other two-stage detectors.

Further, multimodal fusion [20] aims to find the combination of apparently disparate multi-domain data (e.g., visible light, infra-red light, sound, laser) sources to produce a more robust and more rich fused signal. One of the hot topics of the current sensor fusion literature is choosing the level in which the fusion takes place in the detection pipeline [5,21,22]. Sensor fusion happens in three levels, such as: (i) *early fusion* as the pixel-level fusion; (ii) *intermediate fusion* as the feature-level fusion; (iii) *late fusion* as the decision-level fusion. Several studies have been conducted by fusing sensory data in different levels of the perception pipeline. Farahnakian and Heikkonen (2020) [10] achieved state-of-the-art performance to detect marine vessels by applying intermediate fusion techniques using thermal and RGB cameras. However, reasonable real-time inference durations (mostly in the order of seconds) were not achieved. Additionally, the authors even suggested the possibility of adding a LiDAR source to the fused input to explore possible improvements. Liu et al. (2022) [12] proposed an intermediate fusion approach using LiDAR and visual camera data for car detection. Choi and Kim (2021) [11] followed an early fusion approach by combining an infrared camera with a three-dimensional LiDAR sensor, achieving reliable performance in impractical environmental conditions for vision-based sensors. Azam et al. (2019) [23] suggested another early fusion approach that fuses three-dimensional LiDAR with thermal images, ensuring reliable performance for both day and night light conditions. Bhanushali et al. (2020) [24] achieved reliable real-time object detection by training an end-to-end SSD detector on the KITTI dataset [25], merging early and late fusion principles. The former study also suggested the addition of other sensors, such as radar, to increase the robustness of the proposed model.

3. Multimodal Early Fusion Approach for Fiducial Marker Detection

Offshore robotic applications operate in challenging weather conditions where corrupt sensor information or even sensor failure situations are expected. Therefore, robotic solutions need to be resilient and endow redundant and heterogeneous perception systems [13]. Thus, based on the available multimodal sources, this research proposes a simple methodology that produces an expressive and redundant joint representation over a multimodal feature space. The proposed system comprises an (i) early-fusion module that applies a novel early-fusion technique and a (ii) lightweight YOLO-based detector that is fed with an early fused input. A system high-level perspective is depicted by the Figure 1.

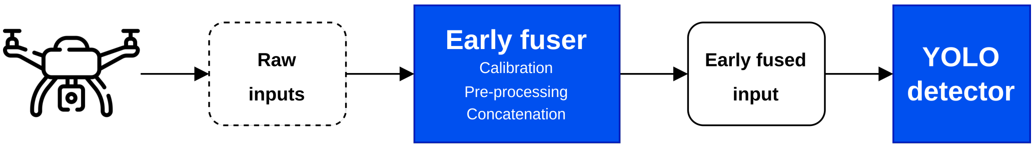


Figure 1. The early-fusion detection system high-level view.

3.1. Early Fuser

Based on spatiotemporal alignment of data streams, the early fuser produces a concatenated 3-channel RGB input containing a joint representation of visual, infrared, and LiDAR modalities. The aim is to anticipate redundancy across input streams and facilitate the detector's feature extraction process. The early fuser procedure is described by: (i) a calibration step where modalities are aligned into a common coordinate system, (ii) a pre-processing step where relevant features for each modality are extracted, and (iii) a final concatenation step where processed modalities are aggregated into a redundant RGB representation.

3.1.1. Calibration

To aggregate sensory information into a common three-channel image, spatial relationships between sensors must be obtained. The visual camera is chosen to be the main frame of reference. Hence, the remaining infrared and LiDAR information are projected into the visual camera image coordinate system. Before projecting the three-dimensional LiDAR point-cloud into the visual image coordinate system, the point-cloud is transformed into the visual camera coordinate system. The three-dimensional spatial relationship between each sensor is described by an extrinsic matrix. Using Zhang's method [26], an extrinsic calibration operation is performed to obtain the extrinsic transformation matrix $E_{l,v} = \begin{bmatrix} R & T \end{bmatrix}$. $E_{l,v}$ is a rigid body transformation described by a rotation

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}$$

and a translation $T = \begin{bmatrix} t_x & t_y & t_z \end{bmatrix}^T$ between the visual camera coordinate system v and the LiDAR coordinate system l, where LiDAR points $P_l = \begin{bmatrix} x_l & y_l & z_l \end{bmatrix}^T$ are converted into visual camera points $P_v = \begin{bmatrix} x_v & y_v & z_v \end{bmatrix}^T$ as follows:

$$\begin{gathered} P_v = E_{l,v} P_v \\ \begin{bmatrix} x_v \\ y_v \\ z_v \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{bmatrix} \begin{bmatrix} x_l \\ y_l \\ z_l \end{bmatrix}. \end{gathered} \tag{1}$$

Following, to transform P_v into pixel points $U_v = \begin{bmatrix} u_v & v_v & 1 \end{bmatrix}^T \in \Re^2$, an intrinsic calibration for the visual camera is performed using the Zhang's method to obtain the intrinsic matrix defined by:

$$K_v = \begin{bmatrix} f_x & \gamma & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix},$$

where (f_x, f_y) are the focal lengths, (c_x, c_y) is the image center and γ is the skew between x and y directions. Therefore, a perspective projection is applied to obtain U_v as follows:

$$\alpha \begin{bmatrix} u_v \\ v_v \\ 1 \end{bmatrix} = K_v \begin{bmatrix} x_v \\ y_v \\ z_v \end{bmatrix},$$

Finally, similar to a stereoscopic camera, an homografy matrix was extracted to map corresponding points between the thermal image and the visual image as carried out by [13].

3.1.2. Pre-Processing

Specific pre-processing techniques are applied for each data stream. The visual stream is undistorted using OpenCV library, to remove natural lens distortion described by K_v. There is no need for rectification because it is a monocular camera. To be represented in a single-channel shape, the visual source, acquired in a RGB representation, is finally converted into a grayscale (single-channel) form. The infrared stream is already acquired in a grayscale representation. Similarly to the visual images, the infrared images undergo undistortion operations by resorting to the OpenCV library using lens distortion described by its intrinsic matrix.

The infrared sensor is properly pre-configured to capture distinct temperature contrast between the ArTuga and the background information. The temperature contrast results in a colour contrast. Availing this natural contrast, a binary threshold filter is applied as an enhancement operation and a filter operation to filter surrounding background noise. The binary operation is a pixel-wise non-linear threshold operation described by:

$$p_{ij}^{binarized} = \begin{cases} 255 & p_{ij}^{orig} > thresh \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where 255 is the maximum color value, $thresh = 128$ is the threshold applied, $p_{i,j}^{orig}$ is any original pixel and $p_{i,j}^{binarized}$ is any transformed pixel. The binary threshold operation is depicted in the Figure 2.

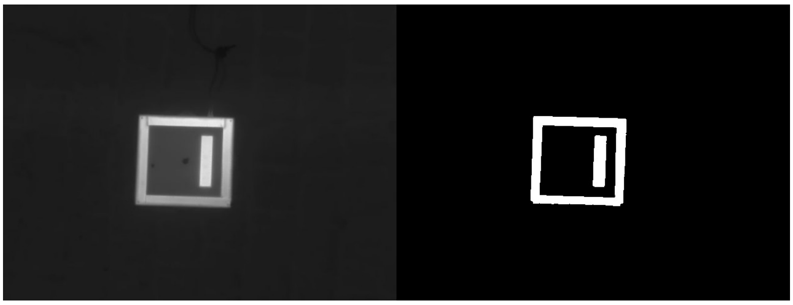


Figure 2. The binary threshold operation applied to the infrared images.

As it can be observed, the binarization results in a more recognizable object by enhancing colour contrast and cleaning the background while preserving edge and corner properties.

The LiDAR stream is acquired in a three-dimensional point-cloud representation. Before intrinsic/extrinsic transformations, this stream is pre-processed to filter and enhance specific inherent information. Excessive and needless point-cloud information is filtered using a voxel downsampling operation. Voxel downsampling is a spatial operation that iteratively buckets points into three-dimensional voxels. Each voxel is compressed and generates a unique three-dimensional point by averaging every inner point. This operation reduces the size of the cloud and throws out excessive information while retaining the overall geometric structure. The used voxels have a leaf size of 5 cm. In addition, more points are removed according to their intensity. The ArTuga has a retro-reflective tape in its borders and white-bit coding area in its interior resulting in disparate values of intensity in comparison to the rest of the point cloud. Points with intensity below a maximum intensity of I are suppressed. Hence, the resulting points mostly belong to the aruco object. A two-dimensional grayscale image is generated by assigning a colour to each resultant pixel, depending on the laser depth, according to a 8-bit scaling operation. The colour c for each pixel is computed as follows:

$$c = 255\left(1 - \frac{d}{m_d}\right),$$

where d is the laser distance measurement and m_d is the maximum depth reached during the UAV flight. A final dilation operation is applied to compensate the sparsity of the image, by applying a 30 × 30 squared kernel as depicted by the Figure 3.

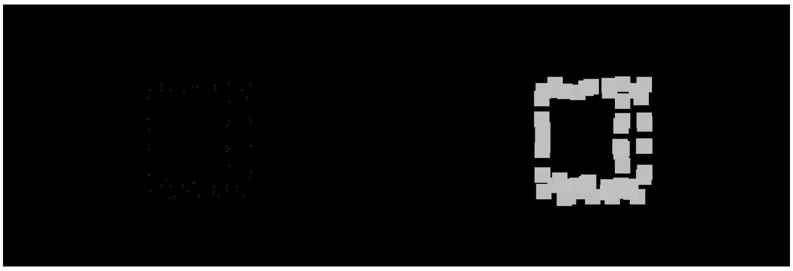
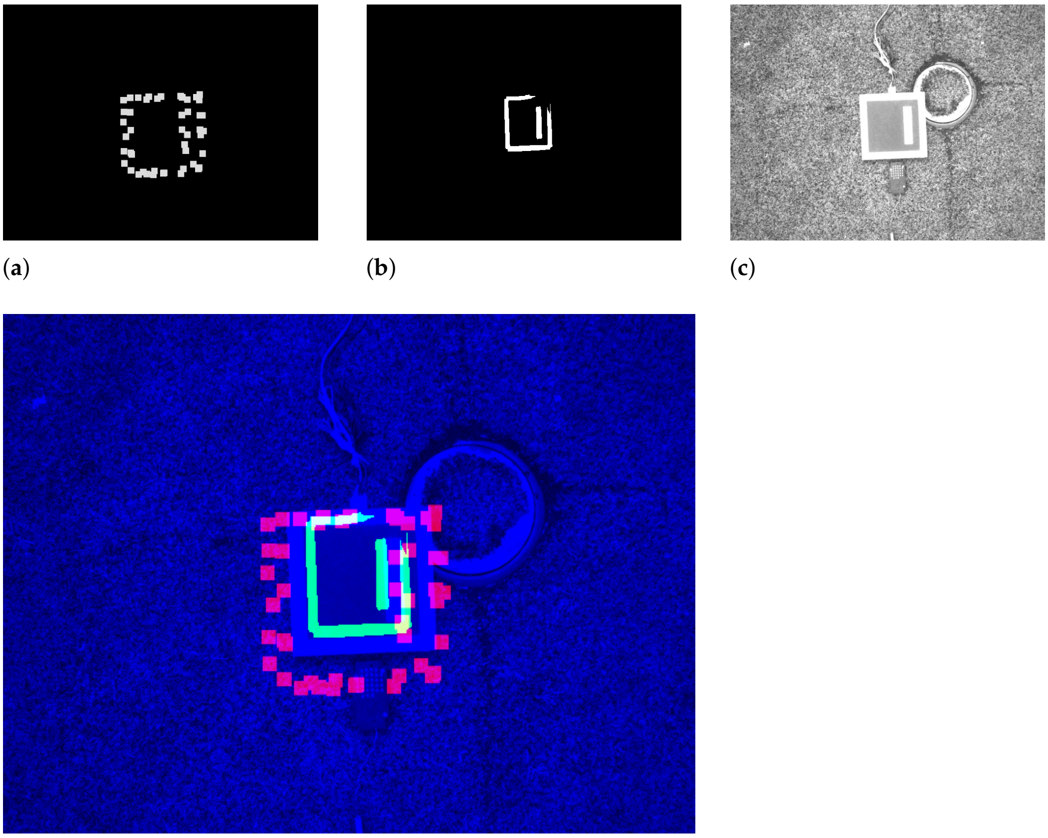


Figure 3. Dilation over the sparsed 2D points.

3.1.3. Concatenation

As a final step, the pre-processed single-channel streams are concatenated into a three-channel RGB image, where the red, green, and blue channels correspond to the LiDAR, infrared, and visual streams, respectively. The channels are presented in the Figure 4.



(**a**)

(**b**)

(**c**)

(**d**)

Figure 4. The individual channels and the aggregated RGB image. (**a**) The LiDAR channel. (**b**) The thermal channel. (**c**) The visual channel. (**d**) The aggregated image.

The concatenation finalizes the early fusion pipeline resulting in an input that is more understandable and correlated across modalities to feed the YOLO detector. The pipeline is depicted in the Figure 5.

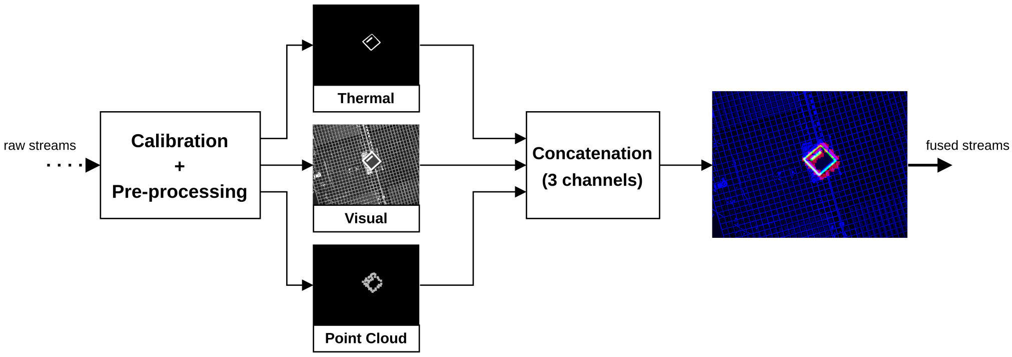


Figure 5. The early fuser pipeline.

3.2. Detector

Nowadays YOLO-based detectors are a solid and popular choice for fast but still accurate detection. Impressively, it is used to tackle real-world tasks (e.g., autonomous cars and UAVs) [27–29], addressing challenging weather conditions [30,31] and real-time speed constraints [32,33]. Nowadays, the most recent versions of YOLO, the YOLOv5 [34], YOLOv6 [35], and the YOLOv7 [36] are the state-of-the-art detectors for fast and accurate detection. This work uses a detector based on the YOLOv7 architecture. Considering real-time demands, the feasibility of the smallest version of the YOLOv7 is explored: the *tiny* version. The aim is to scrutinize the feasibility of these lightweight, faster, and simpler detectors considering the weaker accuracies when compared to the bigger versions. Hence, the proposed early fusion method aims to counteract lightweight limitations by facilitating the input information.

More particularly, YOLO addresses the detection task as a regression problem because it outputs bounding box coordinates (location of objects in the image) and the probability of detection (the confidence in the prediction) for each detected object. YOLO takes a full image as an input, diving it into a into a $n \times n$ grid. For each cell in the grid, it estimates N possible bounding boxes correspondent to N objects. Each detected object is classified by a label l (e.g., $l \in \{pedestrian, truck, traffic\ light\}$). Each bounding box b is defined as a vector as follows:

$$b = \begin{bmatrix} x_{min} & y_{min} & x_{max} & y_{max} & c & l \end{bmatrix},$$

where (x_{min}, y_{min}) and (x_{max}, y_{max}) are the minimum and maximum pixel coordinates, respectively, with respect to the top-left corner of the image, c is the confidence of the prediction containing an object and l is the object class label. Ultimately, YOLO outputs a $N \times 6$ dimensional tensor, where only the most trusted predicted bounding boxes remain, by applying a technique called Non-Maximum-Suppression.

4. Experimental Setup

A UAV named CROW (copter robot for offshore wind-farms), based on a quadcopter frame (with a wingspan of 0.7 m, and a maximum payload of 2 Kg), was remotely operated to execute a few minutes duration flight around a maritime platform that contains the ArTuga marker in its center. The CROW endows a perception system comprised of a (i) three-dimensional LiDAR, a (ii) visual camera, and an (iii) infrared camera. The detection performance is highly dependent on the sensor characteristics. Thus, the sensor choice must provide high resolution capabilities. The set of sensors have the following specifications:

- Visual Camera—*The Imaging Source DFM 37UX273-ML*—Frame Rate: 15 Hz, Resolution: 1440 × 1080 pixels, Field of View: 45° horizontal;

- Thermal Camera—*FLIR Boson 640 Radiometry*—Frame Rate: 15 Hz, Resolution: 640 × 512 pixels, Field of View: 50° horizontal, Temperature Measurement Accuracy: ±5 °C;
- 3D LiDAR—*Ouster OS1-64*—Frame Rate: 10 Hz, Resolution: 64 × 1024 channels, Range: 120 m, Accuracy: ±0.05 m, Field of View: 360° horizontal and 45° vertical.

During the flight, the sensor data is recorded to be further processed to generate the datasets for model training. There was a concern to produce a heterogeneous dataset, therefore the flight operation was conducted considering different spatial perspectives of the platform. Despite the aim of this application being offshore, onshore samples were also acquired to promote heterogeneity. The high variability of an heterogeneous dataset ensures that a model is robust against unexpected and different image perspectives. Naturally, robustness is a major priority to avoid overfitting while training. The CROW UAV executing a landing operation is depicted by the Figure 6.

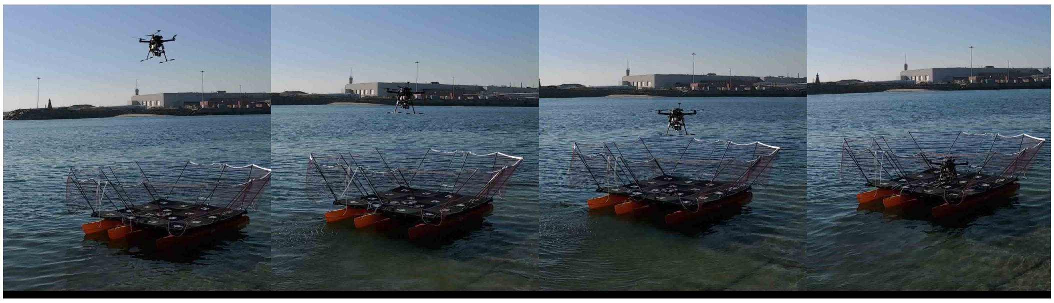


Figure 6. The conducted real experiment of a landing procedure using the CROW UAV.

4.1. Datasets

The performance of the early fusion detector is going to be compared with three unimodal detectors correspondent to each sensor stream. The unimodal detectors function as fine-tuned baseline detectors for each modality. Hence, the aim is to use them as a performance reference for the multimodal early fusion detector. Accordingly four datasets are generated: a visual, a thermal, a LiDAR and a fusion datasets. The datasets are available in a Google Drive public repository [37].

4.1.1. Visual Dataset

The visual dataset comprises 1449 images representing onshore and offshore visual samples from different spatial points of view of the ArTuga with different backgrounds (e.g., landing platform, water, ground). Besides the inherent variability of the data, some data augmentation techniques are applied for this dataset to prevent overfitness. Brightness variation is a clear augmentation technique to train the model against brighter and dark scenarios. The motivation for this technique is to prepare the detector for sunny and night settings. In more depth, for every image, by increasing up to 25% or decreasing up to 90% the brightness, a darker image is a pixel-wise operation, such as $p_{i,j}^{dark} = p_{i,j}^{orig}(1-\delta), \forall i \in [0, h-1] \subset N, j \in [0, w-1] \subset N$ and a brighter image such as $p_{i,j}^{bright} = p_{i,j}^{orig}(1+\beta), \forall i \in [0, h-1] \subset N, j \in [0, w-1] \subset N$, where $\delta \in]0, 0.9], \beta \in]0, 0.25]$, w, h are the image width and height, respectively, $p_{i,j} \in [0, 255] \subset N$ is a pixel in the ith row and jth column of the image, and 0 and 255 are the minimum and maximum colour values. In addition, flipping and rotation augmentation techniques are applied. The images are resized from the acquisition size 1440 × 1080 to a final size of 640 × 480 pixels. The Figure 7 depicts some samples from the visual dataset.

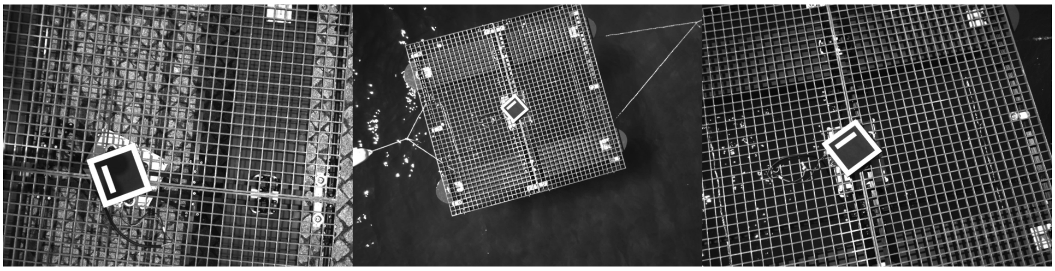


Figure 7. Samples from the visual dataset.

4.1.2. Thermal Dataset

The thermal dataset comprises 441 pre-processed binarized infrared images as already described in the Section 3.1.2. The data augmentation techniques applied for this dataset are flipping and rotation. The images are resized from the acquisition size 1440 × 1080 to a final size of 640 × 480 pixels. The Figure 8 depicts some samples from the thermal dataset.

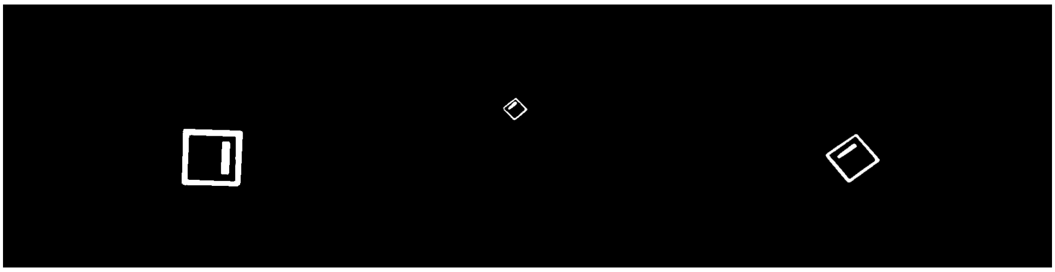


Figure 8. Samples from the thermal dataset.

4.1.3. LiDAR Dataset

The LiDAR dataset comprises 316 pre-processed two-dimensional point cloud projections already described in the Section 3.1.2. The data augmentation techniques applied for this dataset are flipping and rotation. The images are resized from 1440 × 1080 to a final size of 640 × 480 pixels. Figure 9 depicts some samples from the LiDAR dataset.

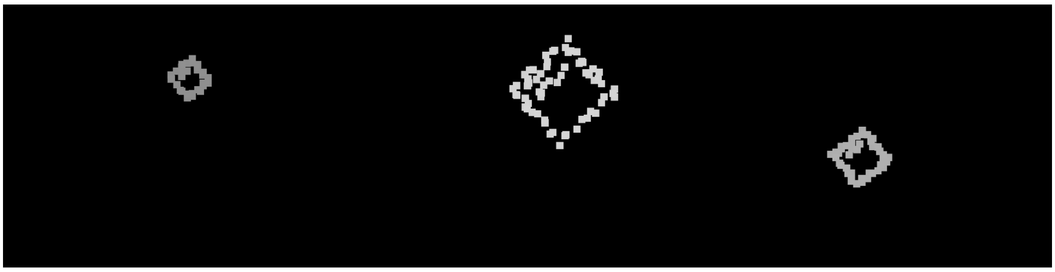


Figure 9. Samples from the LiDAR dataset.

4.1.4. Fusion Dataset

The fusion dataset is constituted by 2158 pre-processed and concatenated RGB images produced by the early fuser described in the Section 3.1. This dataset implements a brightness augmentation technique for the visual channel and flipping and rotation techniques for all channels. The images are resized from 1440 × 1080 to a final size of 640 × 480 pixels. The Figure 10 depicts some samples from the the fusion dataset. As it can be observed, the dataset contains both multimodal samples representing cases when all modalities are available; and unimodal samples representing sensory failure cases when specific modalities are deactivated. Training with unimodal samples intends to promote resilience against sensory failure.

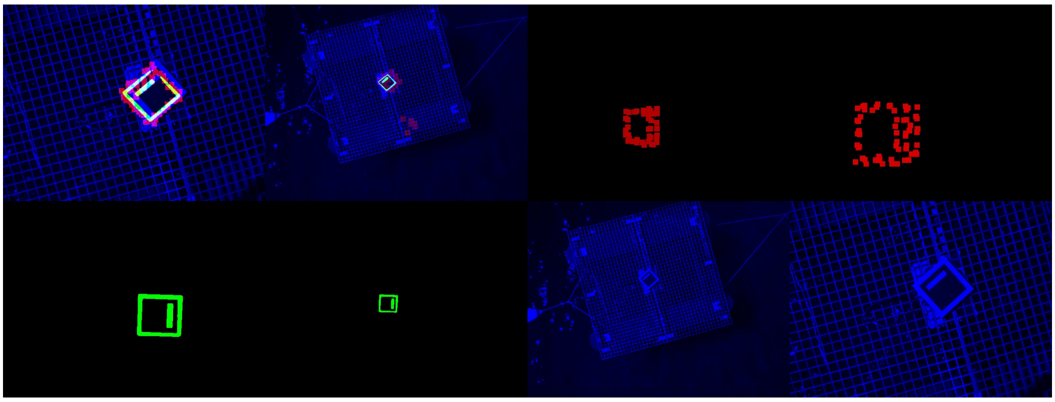
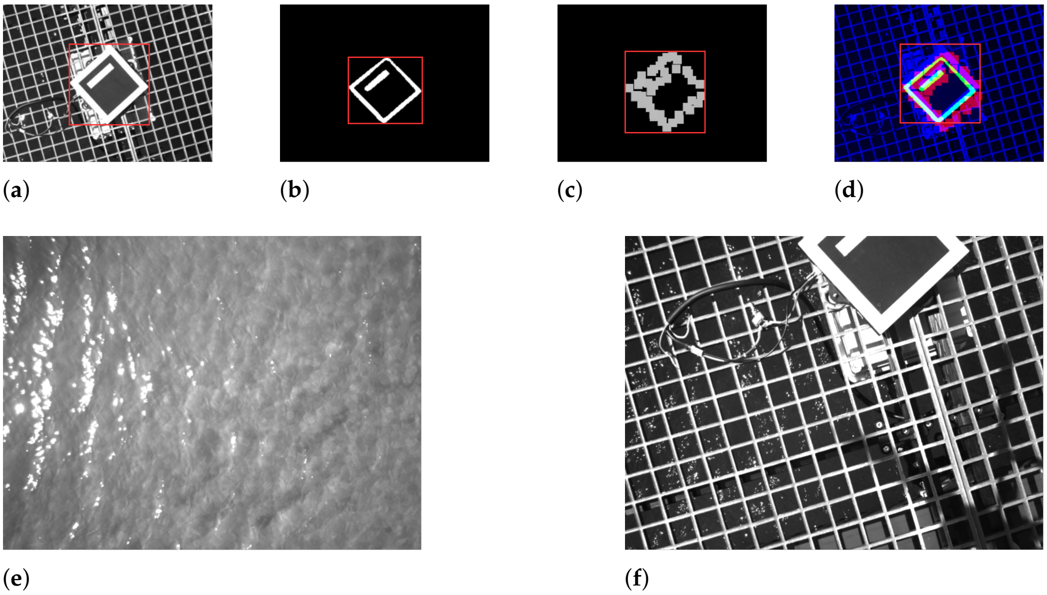


Figure 10. Samples from the Early Fusion dataset.

4.1.5. Annotation

For each sample in the dataset, there is a correspondent annotation. Manual annotations in the form of a bounding box are performed using roboflow framework [38]. For the unimodal datasets, the bounding box fully encloses the unique object present as shown in the Figure 11a–c. In this way, it is noted that there is a care for pixel tightness to drive the model to the best accuracy. As for the case of the fused sample, the annotation considers the stream with the best resolution (shape and colour), as shown in Figure 11d. Additionally, negative samples are present in the dataset to teach the model when an object is not present as shown in the Figure 11e or partially present as depicted by the Figure 11f. For these negative cases, the annotation is not performed, instead it is marked as *null*.



(**a**) (**b**) (**c**) (**d**)

(**e**) (**f**)

Figure 11. Non-null annotations on the top and null annotations on the bottom. (**a**) A visual sample annotation. (**b**) A thermal sample annotation. (**c**) A LiDAR sample annotation. (**d**) A fused sample annotation. (**e**) A null example (no object present). (**f**) A null example (partially present object).

5. Results

This section exposes (i) the training settings and results, (ii) an ablation test to evaluate the model's resilience during landing for all cases of sensor failure, and (iii) a resilience test against challenging weather conditions for fog, dark, and glary scenarios.

5.1. Training

5.1.1. Training Settings

The training hyperparameters were tuned according to useful guidelines towards reliable training performance provided by the official documentation of the YOLO framework (https://github.com/ultralytics/yolov5/wiki/Train-Custom-Data—Accessed on 19 February 2023) . The training procedure is executed for 400 epochs. The selected batch size is 32. The software training platform resorts to Google Colab's servers (https://colab.research.google.com/—Accessed on 19 February 2023), having access to its free of charge NVIDIA Tesla T4 GPUs. An IoU threshold of 0.2 is chosen. A learning rate cosine scheduling [39] is applied with an initial learning rate of 0.01. An Adam optimizer is used because over SGD and RMS Prop it had the best mAP@0.5:0.95 performance. To prevent overfitting, data augmentation techniques inherently exist in the dataset, and a weight decay of 5×10^{-4} is applied. The training hyperparameters are presented in Table 1.

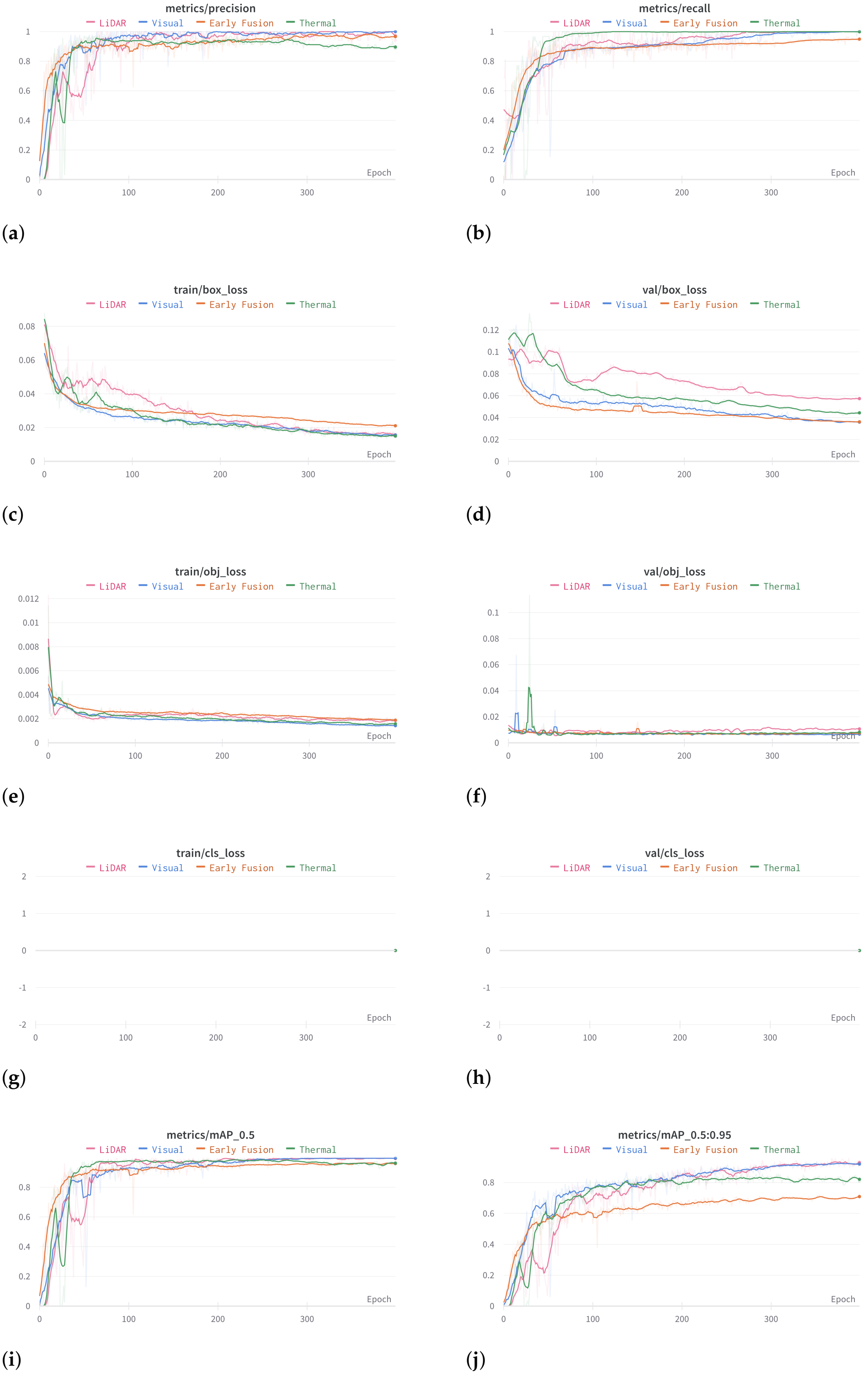
Table 1. The main training hyperparameters.

Parameter	**Quantity**
Epochs	400
Batch size	32
IoU threshold	0.2
Momentum	0.937
Learning rate	1×10^{-2}
Weight decay	5×10^{-4}

5.1.2. Training Results

The training results for all the detectors are presented in the Figure 12. Commonly, a clear training convergence is observed for all models because the train regression loss (`train/box_loss`) and the objectness loss (`train/obj_loss`) monotonically decrease. The *box* loss describes the decrease in the regression error between the predicted and ground-truth bounding boxes (x, y, w, h) values. The *obj* loss describes the error between the confidence the model has on the object presence and the true presence. A *cls* loss has a particular behaviour because this is a single-class problem: when there is a class, it always the class 0. Therefore the problem reduces to predicting the presence of an object and the better enclosing bounding box. In addition, the model does not overfit because the validation losses monotonically decrease and eventually settle into a final value.

Regarding accuracy, all models achieved high levels of accuracy (above 95%) both on the The mAP@0.5 and The mAP@0.5:0.95 metrics (above 70%). The mAP@0.5:0.95 is significantly lower because it is more demanding. The early fusion model has the lower accuracy on the mAP@0.5:0.95 metric because of the redundancy of the predictions. Due to being a multimodal detector there is always some uncertainty about the true location of the object because the modalities are not perfectly aligned. This misalignment between modalities is caused by the temporal asynchronism between sensors. Hence, on average, the accuracy is always lower when compared to the accuracy achieved by unambiguous unimodal detection. The precision and recall metrics also demonstrate the accuracy convergence across all models. The precision describes the validity of true positive prediction on the universe of the selected true positive labels. For this metric, there is a general convergence towards 1, despite the thermal detector. On the thermal samples, due to the non-linearity of the binarization operation, some detail is occasionally lost which negatively influences the prediction.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**) (**j**)

Figure 12. Training results. (**a**) The precision. (**b**) The recall. (**c**) The train regression box loss. (**d**) The validation regression box loss. (**e**) The train regression object loss. (**f**) The validation regression object loss. (**g**) The train regression class loss. (**h**) The validation regression class loss. (**i**) The Mean Average Precision 0.5 accuracy. (**j**) The Mean Average Precision 0.5:0.95 accuracy.

Similarly, for the `recall` metric, it is clear to observe a convergence towards 1, demonstrating that ultimately the predictions are complete: on the universe of all the images in an epoch containing a marker, the model accurately recognized a marker. The early fusion model slightly loses on the recall. This could be caused by the misalignment of some samples. When this misalignment is pronounced, the redundancy is lost and the model produces false negative predictions. This limitation could be avoided by including more unimodal samples on the early fusion dataset, such that the model trains more on unimodal situations (when the redundancy is not present).

Finally, to evaluate the performance of the detectors outside the training domain we compare the mAP score obtained by inferring the correspondent test datasets. Table 2 exposes the mean average precision for each detector.

Table 2. The performance of each detector on its own test set.

Detector	Modality	mAP@.5	mAP@.5:.95	Precision	Recall
Visual	Unimodal	0.999	0.999	1	1
Thermal	Unimodal	1	1	1	1
LiDAR	Unimodal	0.999	0.999	1	1
Early Fusion	Multimodal	0.989	0.989	1	0.985

All detectors achieve outstanding generalizable behaviour. Comparatively, the thermal detector stands out slightly. The thermal information, specifically, discriminates the object better which facilitates detection. Apart from that, thermographic information alone could suffer from particular extreme heat weather conditions (e.g., hot and sunny days). Thus, visual and point cloud information can complement this limitation. Decently, the early fusion detector nearly matches the fine-tuned detectors.

Lastly, the generalization capability of the detector is evaluated by inferring an external dataset called TNO Image Fusion [40] comprised by fused multispectral images as depicted by the Figure 13. Since this dataset does not contain the ArTuga marker, the aim is to examine the resilience of the detector against false positive predictions. Table 3 exposes the number of false positive predictions for five different levels of confidence thresholds across 127 images.

Table 3. False positive evaluation on an external dataset.

Confidence Threshold (%)	False Positives
60	19
70	9
75	4
80	1
90	0

From Table 3, it is clear that below 75% the model starts to become sensitive. However, in a real application if the threshold is set above 80%, the model can be considered reliable. In conclusion, the results from this evaluation are substantial since they demonstrate the resilience of the detector against novel and noisy information.

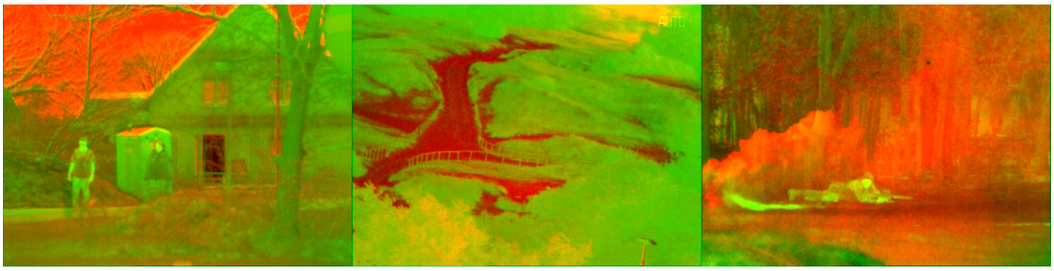


Figure 13. Some images from the TNO dataset [40].

5.2. Testing

The testing phase is comprised by an ablation and a weather tests to evaluate the model's resilience against sensor failure and challenging weather conditions, respectively. Glary, dark, and foggy weather conditions are addressed. Particularly, it is not feasible to operate current UAV technology in rainy conditions, hence rain settings are not addressed. Additionally, it is inopportune to apply LiDAR technology for rainy conditions [41]. The hardware setup for testing comprises an Intel® Core™ i7-10700F CPU @ 2.90GHz × 16 processor and a NVIDIA GeForce RTX 3060 GPU.

5.2.1. Ablation Test

To evaluate the resilience of the proposed early-fusion detector, we perform an ablation study where sensor failure is simulated by intentionally deactivating several combinations of specific modalities during a UAV landing (from an high to a low altitude). Several cases of unable (deactivated) modalities are simulated as follows: only LiDAR, only thermal, only visual, LiDAR and thermal, LiDAR and visual, thermal and visual. As a reference all cases are compared to the baseline (None) where all streams are activated. For each case, 18 images, acquired during landing, are inferred to the model. Recall and inference time results are exposed in the Table 4. True positive predictions are considered for confidences above 80%. For evaluating the accuracy only the recall metric is considered since all the images have a marker present. Moreover, Figure 14 depicts the cases and the predictions.

Table 4. Performance over 18 images during a landing operation for different unable signals.

Unable Signals	True Positives (Conf > 0.8)	Recall	Inference Time (ms)
None	100%	1	5.6
Visual and Thermal	100%	1	5.5
Visual and LiDAR	100%	1	4.2
Thermal and LiDAR	100%	1	4.7
LiDAR	100%	1	4.2
Visual	100%	1	4.1
Thermal	100%	1	4.8

The so needed redundancy against sensor failure is clearly concluded from the results exposed. Solid performance is demonstrated by the Recall results equaling 1 across all unable signal cases. This robustness is expected considering the presence of several unimodal samples in the dataset. Training with unimodal samples prepares the model against unable signals. Summing up, it can be concluded the reliability of the early fusion approach while operating under a sensor failure situation.

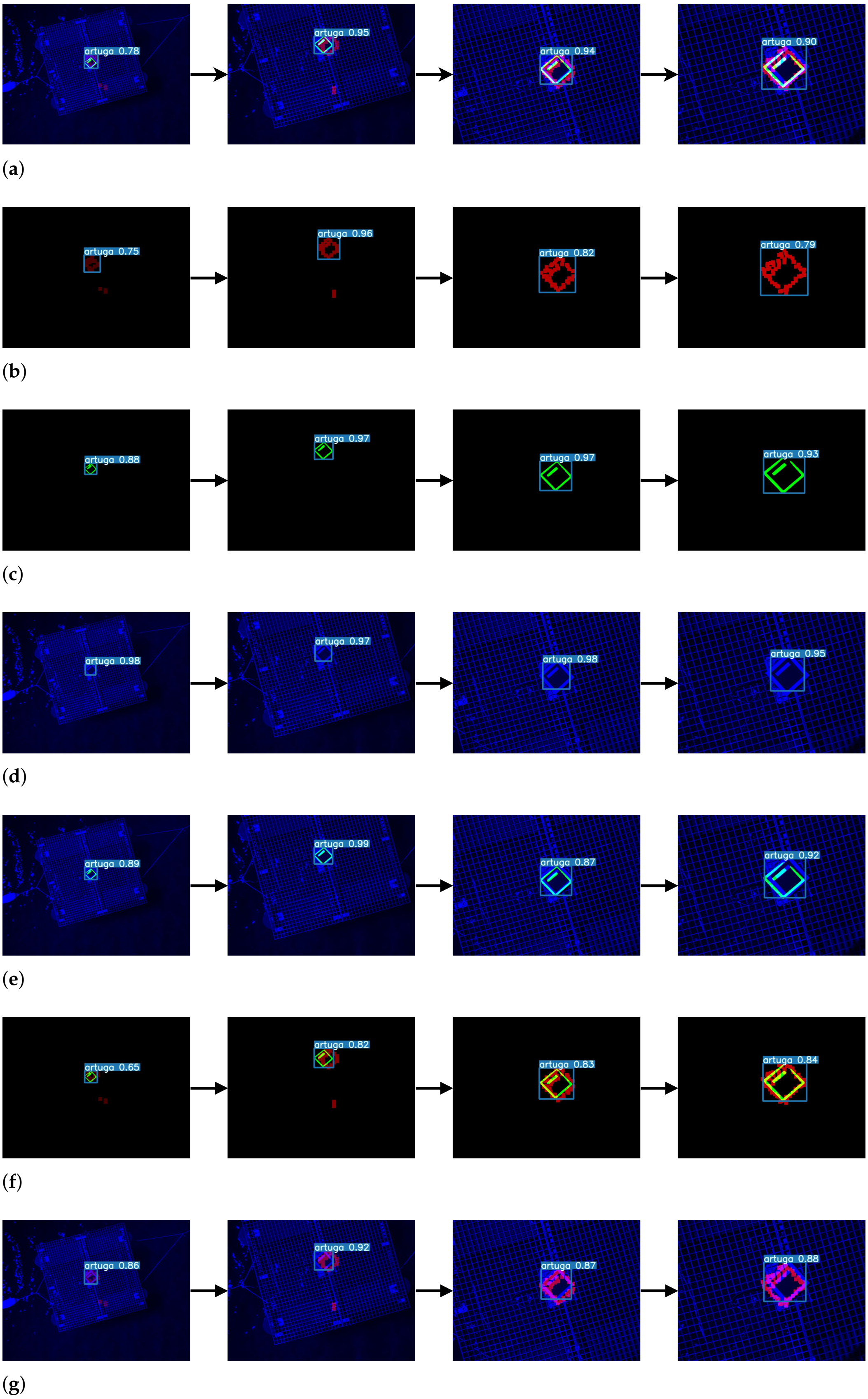
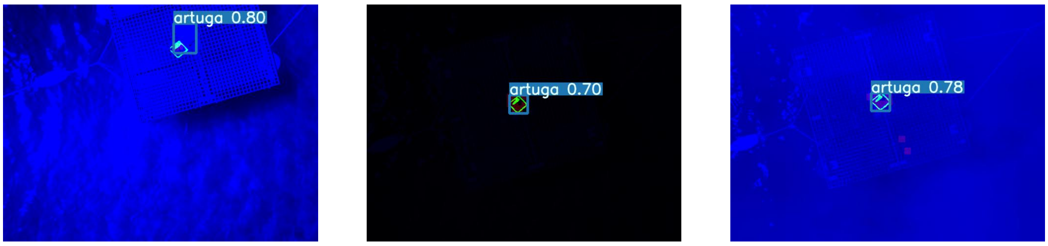


Figure 14. Examples of ablation test predictions for all sensor failure cases. (**a**) Early-fusion (baseline). All sources are activated. (**b**) Relying on LiDAR source only. (**c**) Relying on thermal source only. (**d**) Relying on visual source only. (**e**) LiDAR source failure. (**f**) Visual source failure. (**g**) Thermal source failure.

5.2.2. Weather test

To motivate the use of an early-fusion based detector for challenging weather conditions a stress test for extreme simulated weather conditions is conducted. Glary and dark environments are simulated for increasing and decreasing, respectively, variations in brightness of 10%, 50% and 90% of the original image. Fog environments are simulated by applying stochastic fog augmentation using the Image Augmentation library (https://imgaug.readthedocs.io/en/latest/source/api_augmenters_weather.html—Accessed on 19 February 2023).

Furthermore, every prediction above a 0.5 confidence threshold is considered a true positive. Otherwise, it is considered a false negative. An amount of 100 images are inferred for testing. For every image, there is a marker present, thus, at best, the model should produce 100 true positive predictions. For evaluating the accuracy only the recall metric is considered since all the images have a marker present. The results achieved are presented in the Table 5. Examples of the extreme scenarios are depicted by the Figure 15.



(**a**) (**b**) (**c**)

Figure 15. Examples of predictions under extreme weather conditions. (**a**) A 90% glary (sunny) scenario example. (**b**) A 90% dark (night) scenario example. (**c**) A fog scenario example.

Table 5. Performance over 100 images described by challenging weather conditions.

Weather Condition	True Positives (Conf > 0.6)	False Negatives	Recall	Inference Time (ms)
Bright (10%)	99	1	0.99	2.2
Bright (50%)	98	2	0.98	2.1
Bright (90%)	87	13	0.87	2.0
Dark (10%)	99	1	0.99	2.6
Dark (50%)	100	0	1	2.1
Dark (90%)	96	4	0.96	2.0
Fog (stochastic)	96	4	0.96	2.0

It can be concluded that the model is more sensitive under extreme bright conditions. Conversely, it is resilient against dark and fog conditions. Considering the rare occurrence of extreme brightness situations on the dataset, the model naturally performs poorly against it. Resilience is expected against dark conditions because the dataset includes unimodal samples that mimic sensor failure. Concretely, the model is trained against unable visual signals or, in other words, extreme dark environments. The worthy performance (96% recall) against the stochasticity of fog conditions is the most surprisingly result, which demonstrates the robustness against a more noisy and random signal. Hence, it can be said that the model is presumably robust against real noisy environments.

6. Conclusions

This research introduced a novel methodology that proposes an early-fusion module capable of introducing the required reliability to a lightweight state-of-the-art object detector under sensor failure and extreme weather conditions.

The proposed early fusion approach provided an expressive input that clearly facilitated the detection of a multimodal fiducial marker during a UAV offshore operation. Together, the early fusion detector and the multimodal marker operate in a robust and transparent fashion against challenging weather and sensor failure conditions. In addition, if there is a GPU onboard a robotic solution, it should be emphasized the assurance of a fast system for real-time operating conditions, as demonstrated by the inference time results of less than 6 ms. Consequently, it can be said that assuring robust performance empowered by an early fusion and end-to-end approach for a lightweight detector is one of the major contributions of this work.

Author Contributions: Conceptualization, F.S.N. and A.M.P.; methodology, F.S.N. and A.M.P.; software, F.S.N.; validation, F.S.N., R.M.C. and A.M.P.; investigation, F.S.N.; resources, A.M.P.; data curation, F.S.N. and R.M.C.; writing—original draft preparation, F.S.N.; writing—review and editing, F.S.N., R.M.C. and A.M.P.; visualization, F.S.N., R.M.C. and A.M.P.; supervision, A.M.P.; project administration, A.M.P.; funding acquisition, A.M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871571 (ATLANTIS), and by FLY.PT-P2020 Mobilizador within project POCI-01-0247-FEDER046079. An european patent for the ArTuga multimodal fiducial marker with the reference number EP22212945.4 has been submitted.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data produced by this research is described by four datasets designed to train Early Fusion Multimodal fiducial marker detectors for UAV landing operations. It is available at [37].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pendleton, S.D.; Andersen, H.; Du, X.; Shen, X.; Meghjani, M.; Eng, Y.H.; Rus, D.; Ang, M.H. Perception, planning, control, and coordination for autonomous vehicles. *Machines* **2017**, *5*, 6. [CrossRef]
2. Lim, T.Y.; Ansari, A.; Major, B.; Fontijne, D.; Hamilton, M.; Gowaikar, R.; Subramanian, S. Radar and camera early fusion for vehicle detection in advanced driver assistance systems. In Proceedings of the Machine Learning for Autonomous Driving Workshop at the 33rd Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 14 December 2019; Volume 2.
3. Yeong, D.J.; Velasco-Hernandez, G.; Barry, J.; Walsh, J. Sensor and Sensor Fusion Technology in Autonomous Vehicles: A Review. *Sensors* **2021**, *21*, 2140. [CrossRef] [PubMed]
4. Kim, J.; Koh, J.; Kim, Y.; Choi, J.; Hwang, Y.; Choi, J.W. Robust deep multi-modal learning based on gated information fusion network. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; pp. 90–106.
5. Boulahia, S.Y.; Amamra, A.; Madi, M.R.; Daikh, S. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Mach. Vis. Appl.* **2021**, *32*, 1–18. [CrossRef]
6. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the ICML, Bellevue, WA, USA, 28 June–2 July 2011.
7. Yi, S.; Jiang, G.; Liu, X.; Li, J.; Chen, L. TCPMFNet: An infrared and visible image fusion network with composite auto encoder and transformer–convolutional parallel mixed fusion strategy. *Infrared Phys. Technol.* **2022**, *127*, 104405. [CrossRef]
8. Panigrahy, C.; Seal, A.; Mahato, N.K. Parameter adaptive unit-linking dual-channel PCNN based infrared and visible image fusion. *Neurocomputing* **2022**, *514*, 21–38. [CrossRef]
9. Shopovska, I.; Jovanov, L.; Philips, W. Deep visible and thermal image fusion for enhanced pedestrian visibility. *Sensors* **2019**, *19*, 3727. [CrossRef]
10. Farahnakian, F.; Heikkonen, J. Deep Learning Based Multi-Modal Fusion Architectures for Maritime Vessel Detection. *Remote Sens.* **2020**, *12*, 2509. [CrossRef]
11. Choi, J.D.; Kim, M.Y. *A Sensor Fusion System with Thermal Infrared Camera and LiDAR for Autonomous Vehicles and Deep Learning Based Object Detection*; ICT Express: Jeju, Republic of Korea, 2022.
12. Liu, L.; He, J.; Ren, K.; Xiao, Z.; Hou, Y. A LiDAR and Camera Fusion 3D Object Detection Algorithm. *Information* **2022**, *13*, 169. [CrossRef]
13. Claro, R.; Silva, D.; Pinto, A. ArTuga: A Novel Multimodal Fiducial Marker for Aerial Robotics. *Robot. Auton. Syst.* 2022, *in press*.

14. Soviany, P.; Ionescu, R.T. Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction. In Proceedings of the 2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Timisoara, Romania, 20–23 September 2018; pp. 209–214.
15. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
16. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 1440–1448.
17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv* **2015**, arXiv:1506.01497.
18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. *arXiv* **2015**, arXiv:1506.02640.
19. Liu, W.; Dragomir Anguelov, D.E.; Szegedy, C.; Reed, S.E.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. *arXiv* **2015**, arXiv:1512.02325.
20. Akkus, C.; Chu, L.; Djakovic, V.; Jauch-Walser, S.; Koch, P.; Loss, G.; Marquardt, C.; Moldovan, M.; Sauter, N.; Schneider, M.; et al. Multimodal Deep Learning. *arXiv* **2023**, arXiv:2301.04856.
21. Katsaggelos, A.K.; Bahaadini, S.; Molina, R. Audiovisual Fusion: Challenges and New Approaches. *Proc. IEEE* **2015**, *103*, 1635–1653. [CrossRef]
22. Gadzicki, K.; Khamsehashari, R.; Zetzsche, C. Early vs. Late Fusion in Multimodal Convolutional Neural Networks. In Proceedings of the 2020 IEEE 23rd International Conference on Information Fusion (FUSION), Rustenburg, South Africa, 6–9 July 2020; pp. 1–6. [CrossRef]
23. Azam, S.; Munir, F.; Sheri, A.M.; Ko, Y.; Hussain, I.; Jeon, M. Data fusion of lidar and thermal camera for autonomous driving. In Proceedings of the Applied Industrial Optics: Spectroscopy, Imaging and Metrology, Optical Society of America, Washington, DC, USA, 8–10 July 2019; pp. 2–5.
24. Bhanushali, D.; Relyea, R.; Manghi, K.; Vashist, A.; Hochgraf, C.; Ganguly, A.; Kwasinski, A.; Kuhl, M.; Ptucha, R. LiDAR-Camera Fusion for 3D Object Detection. *Electron. Imaging* **2020**, *2020*, 257. [CrossRef]
25. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
26. Zhang, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [CrossRef]
27. Benjumea, A.; Teeti, I.; Cuzzolin, F.; Bradley, A. YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles. *arXiv* **2021**, arXiv:2112.11798.
28. Murthy, J.S.; Siddesh, G.; Lai, W.C.; Parameshachari, B.; Patil, S.N.; Hemalatha, K. ObjectDetect: A Real-Time Object Detection Framework for Advanced Driver Assistant Systems Using YOLOv5. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 9444360. [CrossRef]
29. Al-Qubaydhi, N.; Alenezi, A.; Alanazi, T.; Senyor, A.; Alanezi, N.; Alotaibi, B.; Alotaibi, M.; Razaque, A.; Abdelhamid, A.A.; Alotaibi, A. Detection of Unauthorized Unmanned Aerial Vehicles Using YOLOv5 and Transfer Learning. *Electronics* **2022**, *11*, 2669. [CrossRef]
30. Kasper-Eulaers, M.; Hahn, N.; Berger, S.; Sebulonsen, T.; Myrland, Ø.; Kummervold, P.E. Detecting heavy goods vehicles in rest areas in winter conditions using YOLOv5. *Algorithms* **2021**, *14*, 114. [CrossRef]
31. Jung, H.K.; Choi, G.S. Improved YOLOv5: Efficient Object Detection Using Drone Images under Various Conditions. *Appl. Sci.* **2022**, *12*, 7255. [CrossRef]
32. Mseddi, W.S.; Sedrine, M.A.; Attia, R. YOLOv5 Based Visual Localization for Autonomous Vehicles. In Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021; pp. 746–750.
33. Liu, W.; Wang, Z.; Zhou, B.; Yang, S.; Gong, Z. Real-time signal light detection based on yolov5 for railway. *IOP Conf. Ser. Earth Environ. Sci.* **2021**, *769*, 042069.
34. Ultralytics. YoloV5. Available online: https://ultralytics.com/yolov5 (accessed on 19 February 2023).
35. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
36. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
37. (2023): Early Fusion Datasets for UAV Landing Operations Using Multimodal Fiducial Marker. Figshare. Dataset. Available online:. (accessed on 19 February 2023). [CrossRef]
38. Roboflow. 2022. Available online: https://roboflow.com/ (accessed on 21 January 2023).
39. He, T.; Zhang, Z.; Zhang, H.; Zhang, Z.; Xie, J.; Li, M. Bag of tricks for image classification with convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 558–567.

40. Toet, A. *TNO Image Fusion Dataset*; Springer Data in Brief: Soesterberg, The Netherlands, 2022.
41. Heinzler, R.; Piewak, F.; Schindler, P.; Stork, W. Cnn-based lidar point cloud de-noising in adverse weather. *IEEE Robot. Autom. Lett.* **2020**, *5*, 2514–2521. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.