applied sciences

MDPI

Article

Exploring a Multimodal Mixture-Of-YOLOs Framework for Advanced Real-Time Object Detection

Jinsoo Kim and Jeongho Cho *

Department of Electrical Engineering, Soonchunhyang University, Asan 31538, Korea; js.kim@sch.ac.kr
* Correspondence: jcho@sch.ac.kr; Tel.: +82-41-530-4960

Received: 24 November 2019; Accepted: 11 January 2020; Published: 15 January 2020

Abstract: To construct a safe and sound autonomous driving system, object detection is essential, and research on fusion of sensors is being actively conducted to increase the detection rate of objects in a dynamic environment in which safety must be secured. Recently, considerable performance improvements in object detection have been achieved with the advent of the convolutional neural network (CNN) structure. In particular, the YOLO (You Only Look Once) architecture, which is suitable for real-time object detection by simultaneously predicting and classifying bounding boxes of objects, is receiving great attention. However, securing the robustness of object detection systems in various environments still remains a challenge. In this paper, we propose a weighted mean-based adaptive object detection strategy that enhances detection performance through convergence of individual object detection results based on an RGB camera and a LiDAR (Light Detection and Ranging) for autonomous driving. The proposed system utilizes the YOLO framework to perform object detection independently based on image data and point cloud data (PCD). Each detection result is united to reduce the number of objects not detected at the decision level by the weighted mean scheme. To evaluate the performance of the proposed object detection system, tests on vehicles and pedestrians were carried out using the KITTI Benchmark Suite. Test results demonstrated that the proposed strategy can achieve detection performance with a higher mean average precision (mAP) for targeted objects than an RGB camera and is also robust against external environmental changes.

Keywords: autonomous driving; RGB camera; LiDAR; CNN; object detection

1. Introduction

Object detection is a fundamental field of computer vision and an essential component of autonomous driving (AD), which uses sensors to detect the driving environment. Sensing the driving environment is closely related to safety, and research on fusion of sensors with different characteristics is being actively conducted to increase the detection rate of objects in a dynamic environment to secure such safety [1]. However, the development of a robust object detection system that can be operated in various driving environments in real time remains a challenge. In recent years, object detection has been considerably improved through in-depth learning algorithms using a large amount of input data. Since the convolutional neural network (CNN) structure was proposed, autonomous driving has seen a large improvement in performance compared to conventional methods. It is also actively used to detect the driving environment [2].

In most cases, various sensors such as RGB cameras, infrared cameras, radars, and LiDARs (Light Detection and Ranging) have been used to build robust object detection systems for the implementation of AD [3]. The detection scheme using an RGB camera basically recognizes the shapes and colors of objects, similar to human vision, and has the highest detection performance when used alone. However, since it is displayed as image data through visible light reflected from an object, it is vulnerable to external environmental factors such as lighting, weather, and object shape [4]. In addition, it is

difficult to obtain accurate three-dimensional distance information of the detected object. To solve such a problem, a method of calculating the distance according to positional difference between pixels using two or more RGB cameras has been proposed, but sensors other than RGB cameras are being used in AD due to the low accuracy of the distance information [5]. Recently, many studies have been performed to overcome the limitations by using LiDAR with RGB cameras to improve object detection [6].

LiDAR emits a laser and represents the signal reflected from objects within the measurement range as point cloud data (PCD). Since the laser derived from the sensor itself measures the reflected signal, it has the advantage of robustness to external environmental factors, unlike RGB cameras that measure visible light. In addition, accurate distance measurement with an object is possible, including reflectance information according to surface properties of the object and distance information according to the reflected time. However, since only the reflected laser signal is measured, a limited image of the surrounding area is generated. As a result, the resolution of the data expressed in PCD is very small, within 10% of the image data, and has a limitation in expressing all the information of the actual environment [7].

As such, RGB cameras and LiDARs have complementary aspects, and there are active proposals for a convergence technology that fuses information from these sensors to enhance object detection performance [8,9]. Recent studies of representative sensor fusions are largely divided into halfway-fusion and late-fusion. The architecture for halfway-fusion newly defines a feature map through fusion of features extracted from each sensor data in the middle of the CNN, and detects objects based on the fused feature map. In the late-fusion structure, a single object detection model is trained based on each sensor data, and the non-maximum suppression (NMS) technique is applied to the result of fusion by concatenating each detection result in a stack at decision level.

In view of halfway-fusion of [10], a method was proposed for detecting a vehicle, by learning an object detection model after fusion of image data and features extracted from an image of a candidate region in which an object may exist in PCD. Recently, a fusion method has been proposed by grouping a single-level feature map obtained by element-wise sum and average of feature maps extracted from image data and PCD and feature maps of each convolution layer [11]. In general, halfway-fusion improves the overall detection performance because feature maps that are fused at the intermediate stages of the CNN contain meaningful information of the object. However, due to the convergence before the decision level, there may be a case where an object cannot be detected by a specific sensor, and thus there is a limit in improving the missed-detection rate.

In the previous study of [12] classified as late-fusion, the object was inferred through a segmentation technique based on image data and PCD, and the outputs classified at decision level were converged based on probability using a convolution-based feature map. This improved the performance of multi-objects classification for vehicles, pedestrians, and cyclists. In addition, a method applying NMS after fusing object detection results was proposed based on three images converted into different sizes according to resolution [13]. Three object detection models were trained in [14] based on image data, LiDAR reflectance, and distance information, and a method of object detection using a multi-layer perceptron (MLP) by extracting features from the detection results has been proposed. Here they implemented late-fusion by redefining reliability through the MLP learning process, which takes the bounding box and its reliability generated by each single object detection model, and targets the ground truth and the intersection of union (IOU) of the bounding box. However, due to the structural nature of MLP, where the number of nodes in the input layer must be constant, the performance of the system was improved only when all the trained detection models detected the objects. In case of failure, the impact of performance improvement through MLP was very limited. In addition, since the bounding box detected during the fusion process is not modified, it is difficult to expect improved detection performance through newly defined reliability when the predicted bounding box and the ground truth have low IOUs.

In this study, we propose an adaptive object detection system that can improve the detection performance by redefining the bounding box through the convergence with multiple sensor detection results even if detection performance of one sensor is degraded by external environmental factors. The proposed system utilizes YOLO (You Only Look Once: Real-Time Object Detection) architecture suitable for real-time object detection and performs independent object detection using image data and PCD. Each of the detection results is then combined to improve the performance of the undetected rate directly at the decision level, late-fusion. In more detail, on the basis of image data and PCD including reflectance and distance information, training for object detection of three CNN-based YOLOs is conducted. The bounding boxes and confidence scores for the objects in each model are then predicted After that, the bounding box is created by applying the weighted average to the coordinates and the sizes of the bounding boxes expected in each model based on the reliability. At this time, to pick a valid bounding box, the bounding boxes predicted by the existing models and the stacking result are applied. Thus, even if all three models fail to detect an object, the bounding box is redefined or the detection result is stacked to compensate for the undetected rate of the system. Furthermore, although all three models detect the object, even if the IOU of the bounding box and ground truth is weak, the system can be strengthened by redefining the bounding box.

For learning and evaluating the performance of the proposed object detection system, object detection was carried out for vehicles and pedestrians using the KITTI Benchmark Suite [15]. The result of sensor fusion using the proposed weighted average shows much better object detection rate than using the RGB camera alone. Even if one YOLO model failed to detect an object, it was possible to reduce the undetected rate by weighting the detection result from the entire model.

2. Object Detection Schemes with RGB Images

2.1. CNN for Object Detection

Object detection in the field of image processing has been performed by extracting a feature of an object from image data in advance and detecting an object based on the feature. To find such a feature point, a scale invariant feature transform (SIFT) architecture extracts local feature points in the image, and a feature descriptor, histogram of oriented gradients (HOG), shows the direction of the edge of the segmented image as a histogram [16,17]. Such an image processing-based method has an inherent disadvantage of the need to find a feature that directly affects object detection performance.

With the advent of CNNs, end-to-end learning of neural networks extracting and learning their own features has led to significant performance improvement in object detection. CNN-based object detection algorithms are largely divided into single-stage and two-stage methods. The single-stage approach typically includes regional CNN (R-CNN), which generates a candidate region of interest (ROI) where objects may exist; extracts features from those regions to detect objects; and regresses the classification algorithms and bounding boxes to detect objects in the ROI. This method shows a large performance improvement compared to the existing image-based object detection algorithms. However, the feature extraction and classification stages are divided, and learning takes a lot of time because the features extracted through input of each ROI to the CNN should be individually trained.

To address these shortcomings, Fast R-CNN and Faster R-CNN with improved learning and detection speed have been proposed. Fast R-CNN reduces learning time by simplifying the computational process and learning steps of CNN through multitasking, which simultaneously learns the loss of classifier and bounding box in the ROI. Faster R-CNN, as shown in Figure 1, uses the region proposal network (RPN) to generate ROI in the last layer of the CNN for faster training and detection speeds. However, the single-stage object detection algorithm performs the two tasks of generating the ROI and classifying the objects in the area sequentially, so that detection accuracy is good, but detection speed is still somewhat slow [18–20].

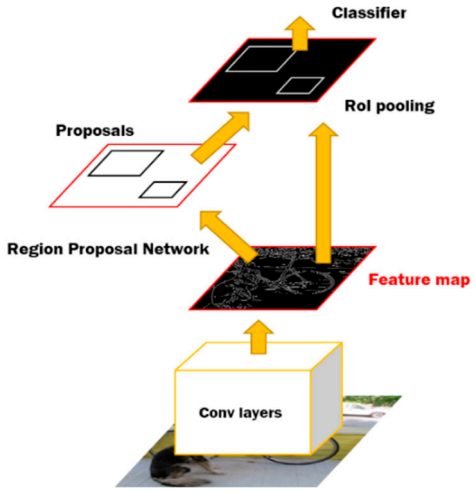


Figure 1. Block diagram of a faster regional convolutional neural network (R-CNN).

2.2. YOLO

YOLO is a state-of-the-art, real-time detection system that predicts and classifies bounding boxes for objects within a given image simultaneously. The image data input to YOLO is divided into S × S grid cells according to resolution, features are extracted through CNN, and the predicted tensor is obtained through a fully connected layer, as shown in Figure 2. The predicted tensor has a size of (S × S) and a length of (B × 5 + C). Here, S × S is the number of grid zones, B is the number of candidate boundary boxes whose center point is included in the grid zone, and C is the number of objects that can be classified. Each lattice zone is represented by a vector having a length of (B × 5 + C), and the set of S × S lattice zones constitutes a predicted tensor of S × S × (B × 5 + C).

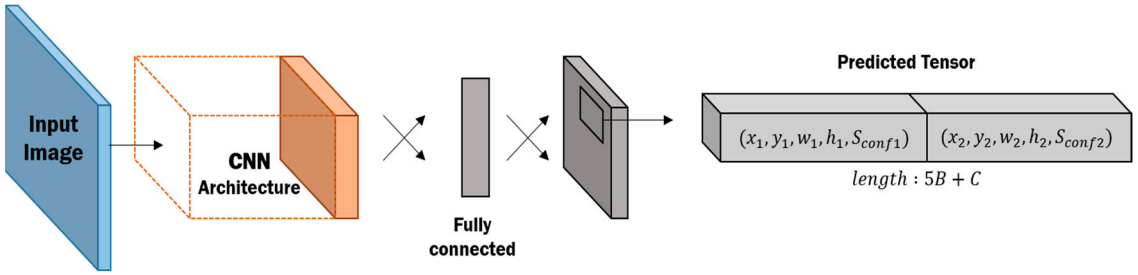


Figure 2. Structure of a You Only Look Once (YOLO) network.

The grid region predicts B bounding boxes, which contain five pieces of information (x, y, w, h, S_{conf}); (x, y) is the center coordinate of the bounding box, (w, h) is the width and height, S_{conf} as in Equation (1) is the product of Pr(Object) and the probability that an object is included in the bounding box, and IOU_{pred}^{truth} is the area relative to the ground truth and intersection of union (IOU), indicating how accurately the bounding box predicted the geometric information of the object.

$$S_{conf} = \Pr(Object) \times IOU_{pred}^{truth} \quad (1)$$

If the actual coordinates and the center coordinates of the predicted bounding box are included in the same grid area, the bounding box is considered to contain the object, and Pr(Object) is calculated as 1; it is treated as 0 if it is included in a different grid area. The IOU is calculated as in Equation (2) by

dividing the intersection of two regions by the union of the two regions and is used to evaluate the accuracy of the predicted bounding box ($b.b_{pred}$) with respect to the ground truth ($b.b_{truth}$).

$$IOU_{pred}^{truth} = \frac{area\left(b.b_{pred} \cap b.b_{truth}\right)}{area\left(b.b_{pred} \cup b.b_{truth}\right)} \tag{2}$$

In addition, grid area is expressed as $\Pr(Object)$ by calculating the conditional probability indicating which among C objects can be classified as the types included in the bounding box as follows:

$$P_{class} = \Pr(Class|Object) \tag{3}$$

In this way, after the tensor of length (B × 5 + C) predicts all the grid regions (S × S), it is extended to obtain the confidence score, CS_{conf}, for classifying objects representing both S_{conf}, which is the probability that the object is included in the predicted bounding box, and P_{class}, indicating the probability that the classified object matches the ground truth, through the following:

$$\begin{aligned} CS &= S_{\text{conf}} \times P_{class} \\ &= \Pr(Object) \times IOU_{pred}^{truth} \times \Pr(Class|Object) \\ &= \Pr(Class) \times IOU_{pred}^{truth} \end{aligned} \tag{4}$$

Finally, for the classified object, the bounding box with the highest CS among the predicted B bounding boxes of the input tensor is selected as the bounding box of the object [21].

3. Multimodal Mixture-Of-YOLOs Framework for Object Detection

The proposed multimodal YOLO-based object detection framework using the weighted mean consists of a data preprocessing unit and a sensor fusion unit. In the preprocessing, the PCD representing the three-dimensional spatial information is projected in two-dimensional space through coordinate correction that matches the viewpoint of the RGB camera. After projection, the depth map and reflectance map are built according to the distance and reflectance information included in the PCD and are used for object detection. The overall schematic diagram of the proposed detection architecture is shown in Figure 3.

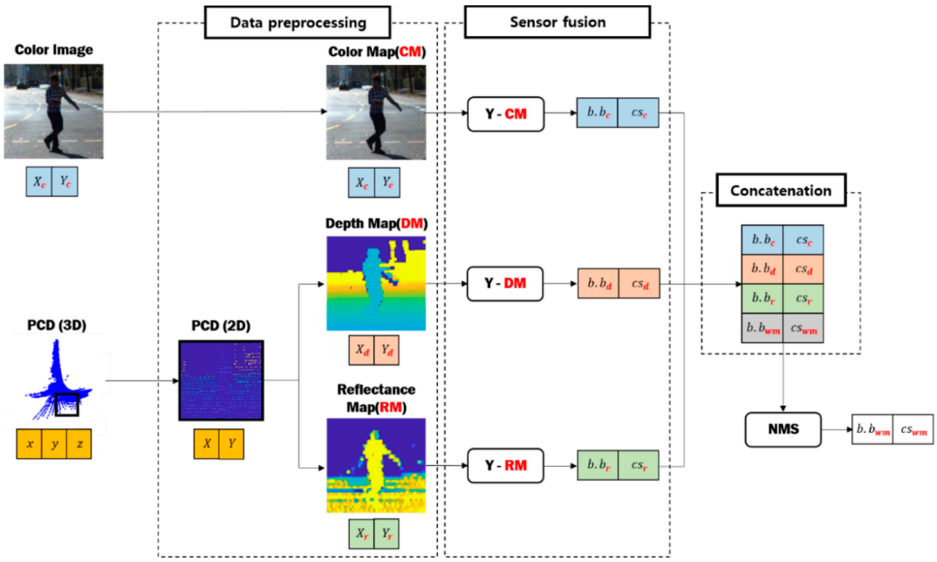


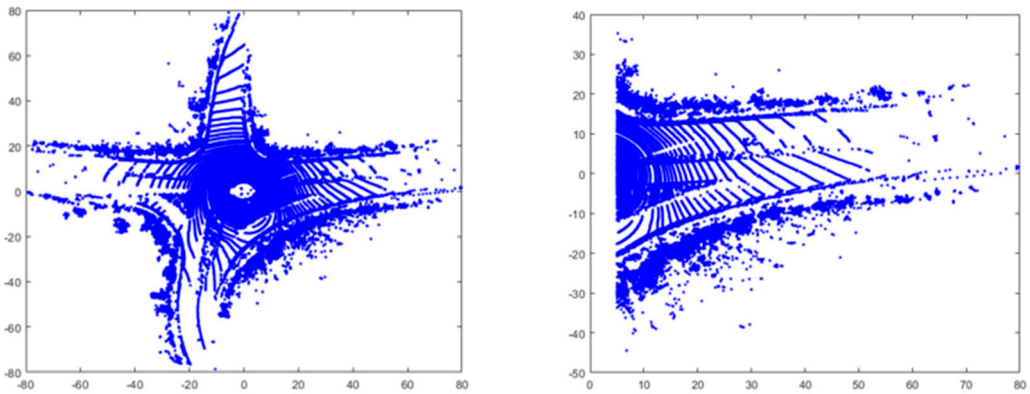
Figure 3. Schematic of the proposed multimodal mixture-of-YOLOs framework.

In the sensor fusion process, each YOLO-based model is trained based on the image data from the RGB camera, the depth map, and the reflectance map of the pre-processed PCD to detect the object. Then, the weighted mean process is applied to determine the coordinates and size of the predicted bounding box. Laser signals derived from a LiDAR have a higher pulse than other sensors to measure long distances. Because they measure information reflected from signals derived from the sensors themselves, they are robust to external environmental factors.

3.1. Data Preprocessing

LiDAR provides the reflected laser signal, $\lambda = (\mathrm{x},\ \mathrm{y},\ \mathrm{z},\ \mathrm{r})$, as a three-dimensional coordinate value (x, y, z) and reflectance information (r), which refers to the intensity of the reflected signal according to roughness, color, and materials of the ground and reflective surface of the object. Object detection using LiDAR is classified based on use the 3D coordinate value as is or by projecting it into 2D space of the top or front view [22]. Object detection using the top view allows easy extraction of the direction and speed of movement of the vehicle, although the computational process of this type of object detection is complicated. On the other hand, object detection using the same front view as the RGB camera and driver view is simpler than object detection using the top view.

In this study, the PCD is converted into a two-dimensional pixel coordinate system such as image data through the conversion process of projecting the PCD to the same front view as the field of view (FOV) of the RGB camera. The pixel coordinate system refers to a 2D reference coordinate system of pixels included in the image data [23]. Since the PCD represents data obtained from all directions of x, as shown in Figure 4a, the PCD expressed in the FOV of the RGB camera is separated as shown in Figure 4b. In Figure 4a, the position of the LiDAR is the origin, the position of the RGB camera is the point shifted by 5 axes from the LiDAR, and the central axis of the FOV is the direction parallel to the y axis, satisfying the condition x > 5. The 3D LiDAR coordinate system, which represents spatial information, is converted to a 2D pixel coordinate system through a rotation/translation transformation matrix provided by the KITTI dataset so that it is projected onto the camera's image plane. As shown in Figure 5a, the coordinate $(\eta = x,\ y,\ z)$, which is a separated PCD, is extracted and converted into the homogeneous coordinate of $(x,\ y,\ z,\ 1)$, and η' projected onto the image plane is obtained through the rotation/translation transformation matrix. Here, the homogeneous coordinate extends N-dimensional coordinate of $(a_1,\ a_2,\ \cdots,\ a_N)$ to $N+1$ dimension with respect to nonzero w, and is represented by $(\mathrm{w}a_1,\ wa_2,\ \cdots,\ wa_N)$. A PCD, η', on the projected image plane is represented by $(\mathrm{w}x',\ \mathrm{w}y',\ \mathrm{w})$, where w represents the distance from the camera and the lidar to the projected PCD, η', depending on the nature of the homogeneous coordinates. Dividing w by two-dimensional coordinates of η' converts it to the pixel coordinate system as $\eta' = (x, y)$ and defines it as $\mathrm{H} = (\mathrm{X},\ \mathrm{Y})$.



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

Figure 4. Example of point cloud data (PCD) from a LiDAR (**a**) PCD (Top View); (**b**) Extracted PCD (Top View).

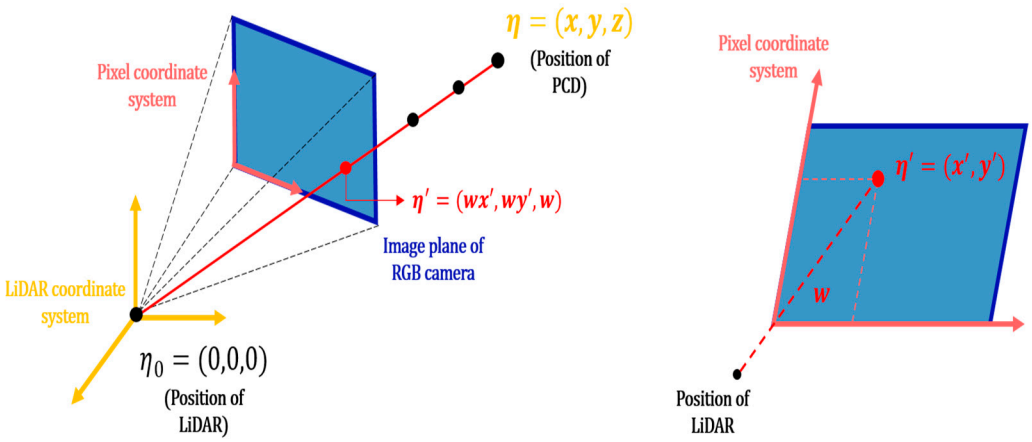


Figure 5. Projection of LiDAR data onto RGB image plane. (**a**) homogeneous coordinate; (**b**) pixel coordinate.

Through this process, the PCD projected by the 2D pixel coordinate system is illustrated in Figure 6a, and Figure 6b shows that the PCD is projected with a view the same as the FOV of the RGB camera. However, since PCD has much lower resolution than the original image, up-sampling for high resolution is performed using spatial interpolation based on bilateral filter [24]. The scaled high-resolution depth map and the reflectance map are generated by applying the weighted pixel values obtained based on the distance and reflectance information of the pixels adjacent to the PCD projected on the image and the adjacent pixels, as shown in Figure 7. The coordinates of the pixels of the depth map and the reflectance map are defined as $H_d = (X_d,\ Y_d)$ and $H_r = (X_r,\ Y_r)$, respectively, and the coordinates of the pixels of the image data are defined as $H_c = (X_c,\ Y_c)$.

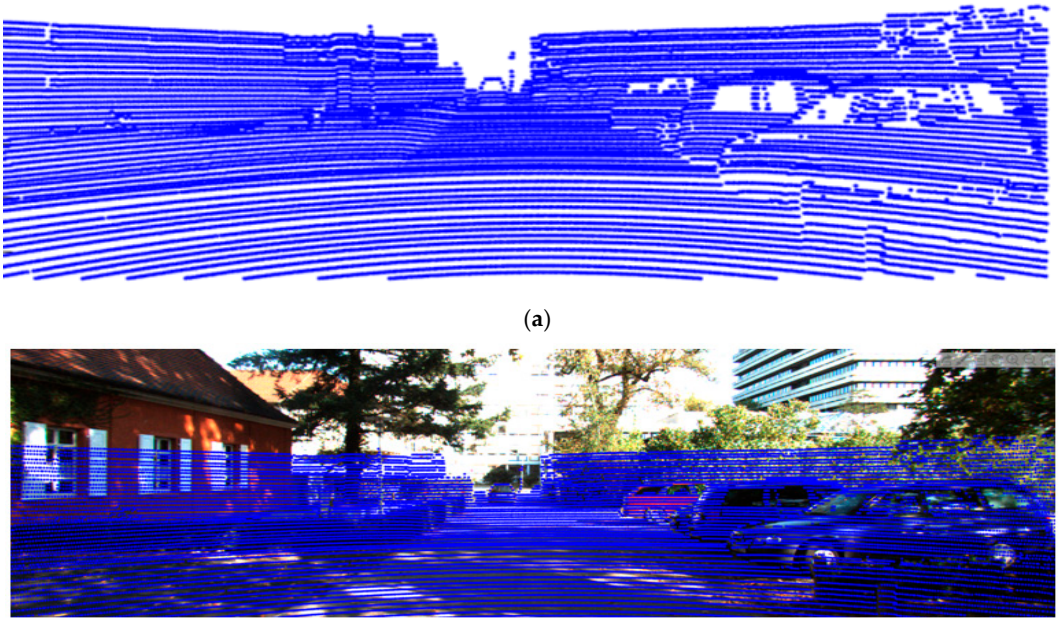


Figure 6. Projection of PCD on an image plane (**a**) PCD converted to pixel coordinates; (**b**) PCD projected on an image.

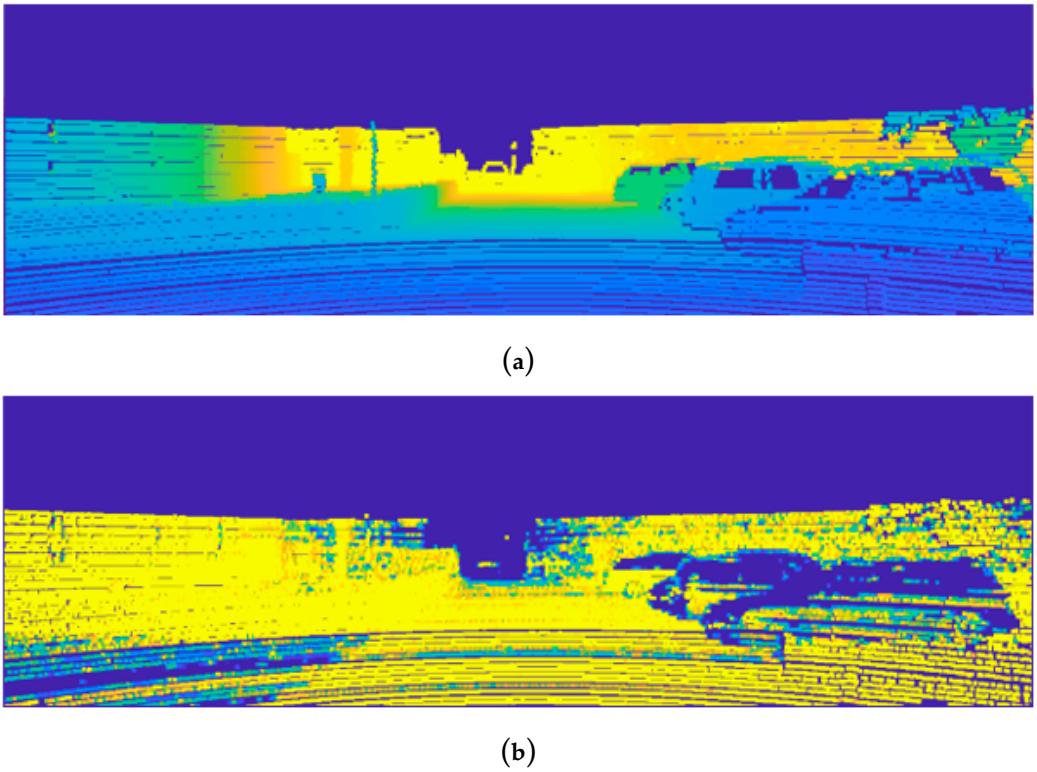


Figure 7. Scaled high-resolution PCD (**a**) depth map; (**b**) reflectance map.

3.2. Object Detection through Multimodal Mixture-Of-YOLO

After preprocessing, each YOLO model marked with Y-CM, Y-DM, and Y-RM is trained and optimized based on the color map, depth map, and reflectance map, respectively, so object detection proceeds independently. Each trained YOLO object detection model outputs the information of the bounding box, $b.b_k = (x_k,\ y_k,\ w_k,\ h_k)$, indicating the position and size of the object included in the data, and CS_k, where $k = c,\ d,\ r$, indicates the detection result as a probability. CS_k reflects the probability that an object is classified, and the bounding box, $b.b_k$, from the object detection result with high CS_k has a high IOU because the area overlapping the bounding box of the ground truth is widened. In general, when there is more than one object, many object detection algorithms apply a non-maximum suppression technique that excludes all bounding boxes other than that with the highest confidence score. Rather, a weighted mean is employed to extract a more accurate bounding box, which results in better convergence of detection results. The element-wise mean of geometric information of the bounding box is obtained by weighting CS_k of the detected objects from the three models as described in Equation (5)

$$b.b_f = \left(\frac{\sum_k x_k CS_k}{\sum_k CS_k},\ \frac{\sum_k y_k CS_k}{\sum_k CS_k},\ \frac{\sum_k w_k CS_k}{\sum_k CS_k},\ \frac{\sum_k h_k CS_k}{\sum_k CS_k} \right) \tag{5}$$

When simply averaging the three bounding boxes without applying weights, they are fused based on the geometric information of each regardless of the ground truth. However, when the average of three bounding boxes is calculated using CS_k as a weight, the bounding box with a high IOU can be obtained because the IOU considers the ground truth. In addition, even if at least one of the three models of Y-CM, Y-DM, and Y-RM fails to detect an object, detection performance can be compensated by a weighted mean based on detection results from the other two models. In this way, the proposed object detection system composed of three YOLO models (Y-CM, Y-DM, Y-RM) and a weighted mean scheme is called a mixture-of-YOLOs with weighted mean defined by MYs-WM.

4. Experimental Results

The KITTI dataset used for performance evaluation in this paper was extracted from an urban area using a vehicle equipped with sensors such as RGB cameras and Velodyne LiDARs and consists of 7481 sequences of training data. The training data include nine types of objects and 51,867 labels, of which 65% (4860) were used for training, 35% (2621) were used for performance evaluation, and the objects were selected as cars and pedestrians. The OS of the workstation used for learning YOLO-v3 [25] was Ubuntu 16.04.5 (4.15.0–38 kernel), and the GPUs were two GTX 1080 Tis (11GB). All parameters except the input data size of YOLO-v3 were used as default values provided by YOLO-v3. Since the resolution provided by the KITTI data set was 1242 × 375, the input data size of YOLO-v3 was modified from the default value of 416 × 416 to match the image resolution of the KITTI data set. The number of training epochs was set to 8000 for each YOLO model, and the performance evaluation of object detection was conducted based on average precision (AP), which is a performance evaluation indicator based on the PASCAL VOC IOU metric and undetected rate. AP is an evaluation index that takes into account both missed detection and false alarm rates and is defined by precision and recall, represented by Equation (6)

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} = \frac{TP}{all\ detections} \\ Recall &= \frac{TP}{TP+FN} = \frac{TP}{all\ ground\ truths} \end{aligned} \tag{6}$$

Here, correct detection of an object is a true positive (TP), while incorrect detection is a false negative (FN). If an object that should not be detected is not detected, it is considered a true negative (TN), and if an object is detected that should not be, it is considered a false positive (FP). Precision is the ratio of what the model classifies as true to what is actually true, and recall is the ratio of what is actually true to what your model predicts to be true.

Precision and recall are affected by IOU, and the numerical value representing the product (the area of the curve) of the increase in recall relative to precision according to IOU is defined as AP. For performance evaluation of the proposed object detection system, we set IOU to 0.7 for cars and to 0.5 for pedestrians according to the KITTI dataset. As mentioned earlier, the proposed system aims to enhance the performance of object detection by fusing all object detection results from Y-CM, Y-DM, and Y-RM through a weighted mean. Performance comparisons with the single-object detection systems (Y-CM, Y-DM, Y-RM) and [1], where Faster R-CNN is applied to VGG-16 [26] structure based on a RGB camera and a LiDAR, were conducted with IOUs of 0.3, 0.5, and 0.7, and the evaluation results are summarized in Table 1 for cars and Table 2 for pedestrians. The results show that Y-CM had the highest detection performance of 87.12% (IOU = 0.7) for cars and 76.62% (IOU = 0.5) for pedestrians among the single-object detection systems, while Y-DM and Y-RM each showed about 16% lower detection performance. Although Y-DM and Y-RM have become high-resolution images through data preprocessing, the detection performance of Y-CM remains the highest since the resolutions of the depth and reflectance maps are less than 35–45% of those of the color map. However, since Y-CM is vulnerable to external environmental changes, detection performance of Y-DM and Y-RM may be better when the background is darkened by shadow or when a part of the object is obscured. Therefore, when the detection results of the single-object detection systems are different from each other, the performance can be improved through the proposed MYs-WM by reinforcement through result fusion. The detection result of the car through MYs-WM was improved to 89.83% (IOU = 0.7), and that of the pedestrian to 79.25% (IOU = 0.5), which is higher than that of the Faster R-CNN-based convergence system of [1]. Examples of fusion detection results of MYs-WM in comparison with Y-CM are shown in Figure 8 for cars and Figure 9 for pedestrians. In the figures, the white dotted line indicates the ground truth, and the red and blue solid lines indicate the bounding box of the object detected by MYs-WM and Y-CM, respectively. In particular, in Figure 9, for the white bounding box representing the ground truth of pedestrians, Y-CM hardly detected pedestrians due to the shadow effect, while MYs-WM detected almost all of them.

Table 1. Comparison of detection of cars according to intersection of union (IOU).

	AP [%]		
Detection Architecture	**IOU = 0.3**	**IOU = 0.5**	**IOU = 0.7**
Y-CM	91.25	90.80	87.12
Y-DM	82.99	81.84	70.57
Y-RM	83.88	82.19	68.59
MYs-WM (ours)	93.64	93.53	89.83
MLF_HHA_Faster-RCNN [1]	-	-	87.90

Table 2. Comparison of detection of pedestrians according to IOU.

	AP [%]		
Detection Architecture	**IOU = 0.3**	**IOU = 0.5**	**IOU = 0.7**
Y-CM	80.53	76.62	53.68
Y-DM	68.49	60.48	27.94
Y-RM	65.22	59.93	29.92
MYs-WM (ours)	81.57	79.25	56.24
MLF_HHA_Faster-RCNN [1]	-	71.40	-

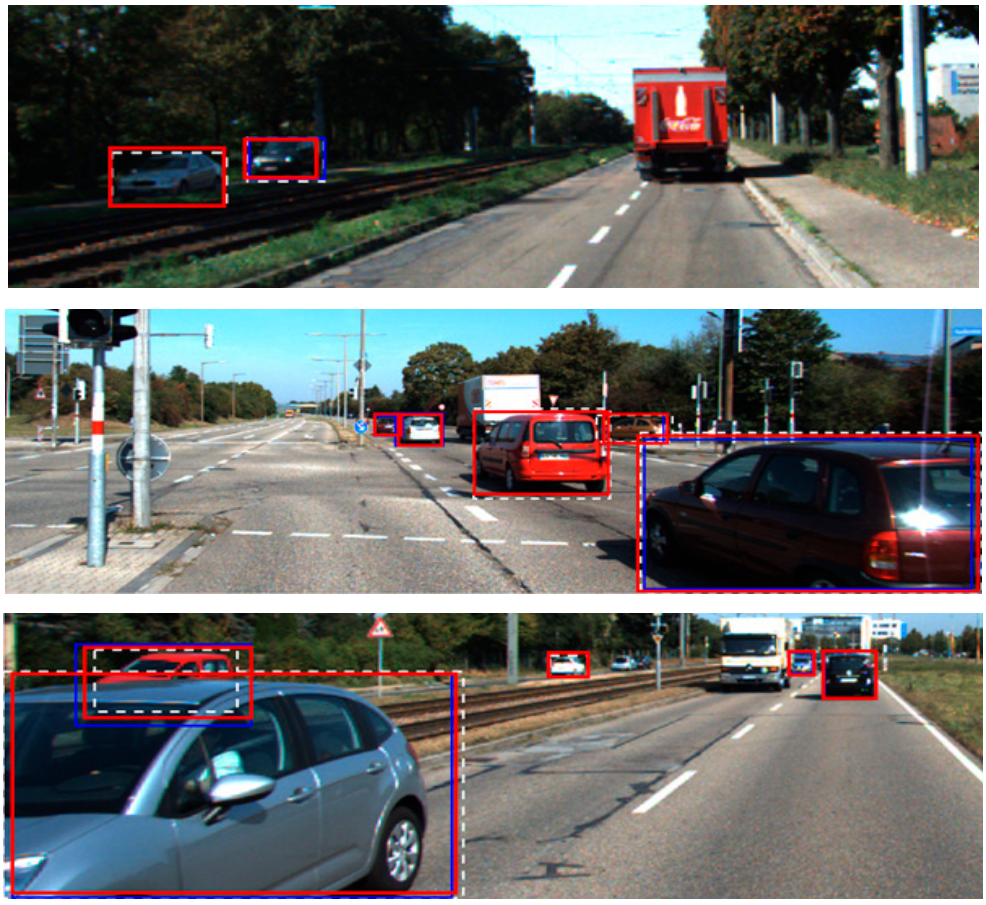


Figure 8. Comparison of car detection examples with Y-CM and MYs-WM.

The performance evaluation method for the KITTI dataset is generally divided into three difficulty levels of 'easy', 'moderate', and 'hard', depending on the size of the object to be detected and the degree of truncation. 'Easy' means 'fully visible' with at least 40 pixels, 'moderate' is 'partial occlusion' with at least 25 pixels, 'hard' is 'higher occlusion', and the number of pixels is equal to 'moderate' [27]. Therefore, an additional performance evaluation was carried out based on AP and missed-detection rate according to the difficulty level of pedestrians and cars, and the detection results are presented in Tables 3 and 4. Additional comparison and evaluation with [28], a system of detecting objects by input

of the left and right images from a stereo camera, its difference map, and the extracted features from PCD into the deformable part models were also performed.

For the detection of cars and pedestrians, the proposed MYs-WM showed the highest performance at all difficulty levels; in particular, MYs-WM showed much higher AP compared to [28] at difficulty levels of 'easy' and 'moderate'. In addition, Figure 10 demonstrates the effectively improved missed-detection rate of the proposed system. Improvement in detection performance is not as high with the new method at the level of 'hard.' This is because single-step object detection algorithms, such as YOLO or SSD, are relatively fast in real time compared to other object detection algorithms due to spatial constraints but are limited in detecting small objects rated as 'hard.' The CNN-based object detection algorithm extracts low-level features from the input image through the convolution layer and then classifies the objects through fully connected nodes. The smaller is an object, the more information about features extracted through the convolutional layer is lost. In particular, YOLO detects images divided into lattice regions using a single regression method, so the detection performance for small objects or multiple overlapping objects is lower than that of the two-stage object detection algorithm of the R-CNN series.

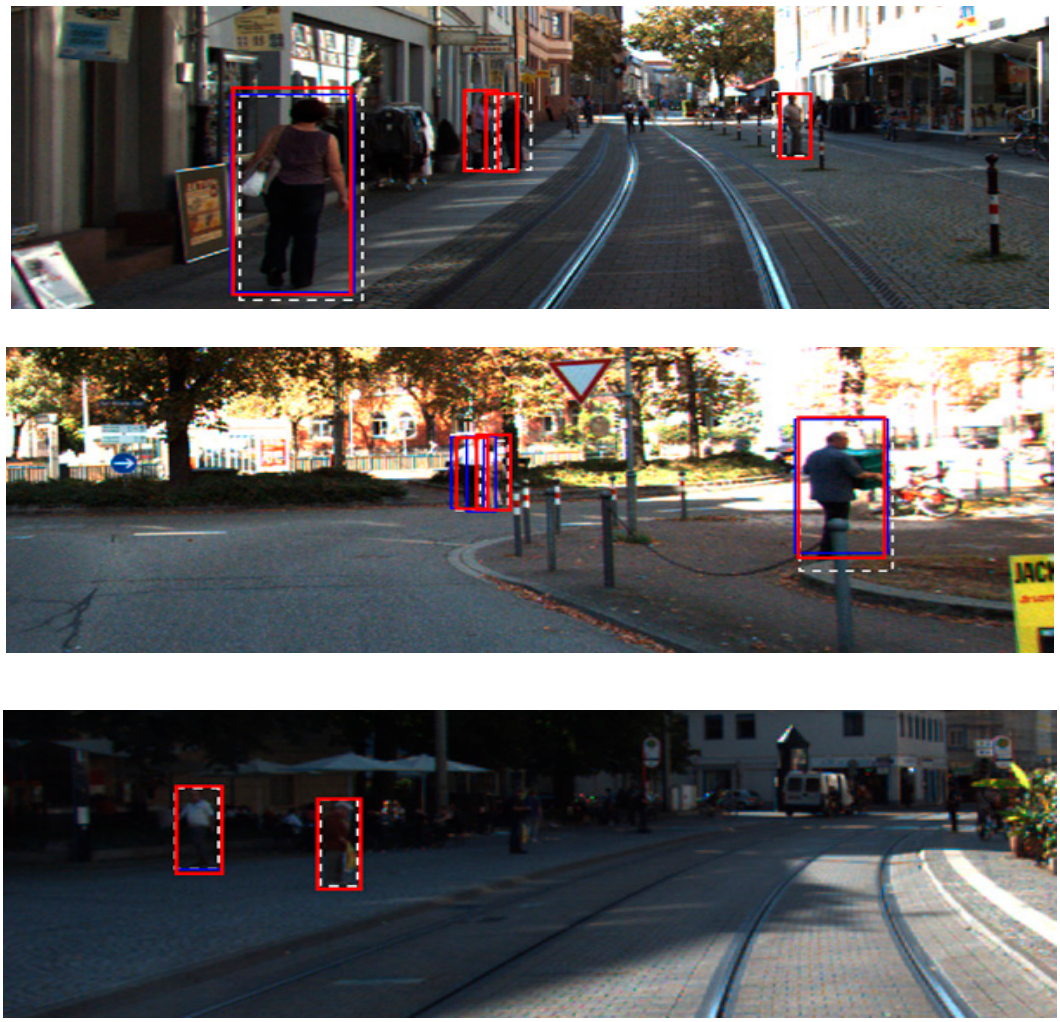


Figure 9. Comparison of pedestrian detection examples with Y-CM and MYs-WM.

With the exception of the 'hard' level, MYs-WM showed relatively high detection performance compared to other detection systems. This is because the detection rate is improved by reinforcing the detection result even though the objects to be detected differ according to the characteristics of the sensor used in the proposed fusion system. Since RGB cameras express data through high-resolution

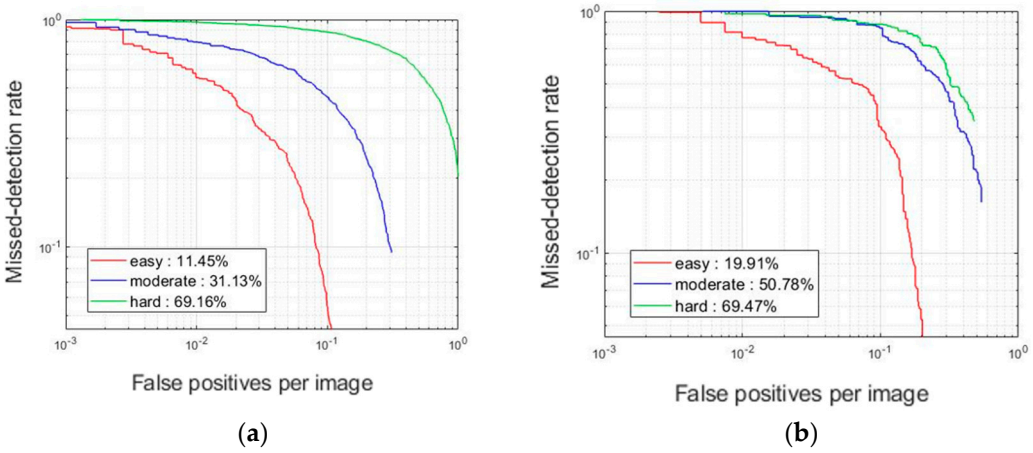
pixels with a wide range of values from 0 to 255, the shape and color of objects are clearly expressed, and detection performance is high. However, detection performance is degraded when affected by external environmental factors such as lighting, shadows, and obstacles. On the other hand, a LiDAR uses a high-pulse laser derived from the sensor itself to acquire the reflected data and is robust to external environmental factors; however, its low resolution results in low object detection performance under normal circumstances.

Table 3. Comparison of detection results of cars classified by difficulty.

	AP [%]		
Detection Architecture	**Easy**	**Moderate**	**Hard**
Y-CM	89.02	74.29	47.75
Y-DM	62.62	50.27	36.49
Y-RM	56.68	41.74	32.06
MYs-WM (ours)	93.74	84.15	54.17
DPM-C8B1 [16]	74.33	60.99	47.16

Table 4. Comparison of detection results of pedestrians classified by difficulty.

	AP [%]		
Detection Architecture	**Easy**	**Moderate**	**Hard**
Y-CM	86.99	61.01	42.75
Y-DM	66.69	40.44	26.34
Y-RM	69.37	35.83	22.04
MYs-WM (ours)	89.25	61.83	43.77
DPM-C8B1 [16]	47.74	39.36	35.95



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

Figure 10. Missed-detection rate by MYs-WM according to difficulty. (**a**) car; (**b**) pedestrian.

Lastly, to evaluate the performance of RGB cameras that are vulnerable to changes in the external environment, the intensity of the image data applied to Y-CM is changed, and various environmental changes are simulated by adding Gaussian white noise. Bright contrast images are used to simulate lightning strikes or high beams, dark contrast images describe the interior of a tunnel or nighttime, and Gaussian white noise represents various external environments, such as snowy, rainy, or foggy weather. We added a constant 50 to decrease the contrast of the image, subtracted the same number to increase it, and generated Gaussian white noise with a zero mean and variance of 0.005 as typical noise in real environments.

Table 5 compares Y-CM and MYs-WM for each external environmental factor in terms of AP. It shows the detection results of Y-CM, which is greatly influenced by external environment change, and those of MYs-WM, which combines the detection results of Y-CM with those of Y-DM and Y-RM and is not significantly affected by external changes. Although the object detection result through the RGB camera is adversely affected by external environmental factors, the weighted mean of the

object detection results through LiDAR can obtain much improved object detection results compared to Y-CM.

Table 5. Robustness to external environment changes.

AP [%]						
Environment Change	**Bright**		**Dark**		**Noise**	
Detection Architecture	**Y-CM**	**MYs-WM**	**Y-CM**	**MYs-WM**	**Y-CM**	**MYs-WM**
Car	84.53	87.71	61.83	82.38	78.11	89.88
Pedestrian	62.96	77.68	55.42	72.67	56.78	78.05

5. Concluding Remarks

In this paper, we proposed a weighted mean-based adaptive object detection strategy that enhances detection performance through convergence of object detection results based on an RGB camera and a LiDAR required for autonomous driving. When the image data of an RGB camera are obtained and a depth map and a reflectance map are generated by the PCD of a LiDAR scaled to high resolution, object detection is performed using the respective Y-CM, Y-DM, and Y-RM models. The object detection results are then combined to demonstrate the final detection performance by the weighted mean scheme. After that, we refined the detection performance by fusing the redefined bounding box by weighted mean of the bounding boxes at decision-level based on the bounding boxes detected by each YOLO-based model and the confidence score. This makes it possible to compensate for the non-detection that occurs when either model fails to predict the bounding box. In addition, even though each single object detection model predicted the bounding box, even if the IOU with the ground truth was low, the detection rate of the system could be improved by redefining the bounding box through the weighted average based on the reliability. As a result, it was verified that the weak point of late-fusion caused by the fusion of each detection result at decision level could be improved through the proposed method. In such a way, object detection performance is improved at an appropriate processing speed in real time by reinforcing the object detection through image data having high resolution but vulnerable to external environmental factors and the detection through PCD having low resolution but strong external environmental changes. Specifically, object detection tests on cars and pedestrians in consideration of the influence of external environment change in the actual driving environment were greatly improved by the proposed MYs-WM. In the future, we plan to continue to supplement the system performance at the decision level by further training and efficiently extracting the 3D predictive tensor based on stack-based learning or reliability scores and the geometric information of the bounding box based on problem analysis results.

Author Contributions: All authors took part in the discussion of the work described in this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MOE) (No. 2018R1D1A3B07041729) and the Soonchunhyang University Research Fund.

Acknowledgments: The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

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

References

1. Banerjee, K.; Notz, D.; Windelen, J.; Gavarraju, S.; He, M. Online Camera LiDAR Fusion and Object Detection on Hybrid Data for Autonomous Driving. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 1632–1638.

2. Yang, M.; Wang, S.; Bakita, J.; Vu, T.; Smith, F.D.; Anderson, J.H.; Frahm, J.M. Re-thinking CNN Frameworks for Time-Sensitive Autonomous-Driving Applications: Addressing an Industrial Challenge. In Proceedings of the 2019 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS), Montreal, QC, Canada, 16–18 April 2019; pp. 305–317.
3. Fritsche, P.; Zeise, B.; Hemme, P.; Wagner, B. Fusion of radar, LiDAR and thermal information for hazard detection in low visibility environments. In Proceedings of the 2017 IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR), Shanghai, China, 11–13 October 2017; pp. 96–101.
4. Kim, J.; Choi, J.; Kim, Y.; Koh, J.; Chung, C.C.; Choi, J.W. Robust Camera Lidar Sensor Fusion Via Deep Gated Information Fusion Network. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 1620–1625.
5. Kondermann, D.; Nair, R.; Honauer, K.; Krispin, K.; Andrulis, J.; Brock, A.; Gussefeldm, B.; Rahimimoghaddamm, M.; Hofmann, S.; Brennerm, C.; et al. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 19–28.
6. Kocić, J.; Jovičić, N.; Drndarević, V. Sensors and sensor fusion in autonomous vehicles. In Proceedings of the 2018 26th Telecommunications Forum (TELFOR), Belgrade, Serbia, 20–21 November 2018; pp. 420–425.
7. Premebida, C.; Garrote, L.; Asvadi, A.; Ribeiro, A.P.; Nunes, U. High-resolution lidar-based depth mapping using bilateral filter. In Proceedings of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil, 1–4 November 2016; pp. 2469–2474.
8. Chavez-Garcia, R.O.; Aycard, O. Multiple sensor fusion and classification for moving object detection and tracking. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 525–534. [CrossRef]
9. Xue, J.R.; Wang, D.; Du, S.Y.; Cui, D.X.; Huang, Y.; Zheng, N.N. A vision-centered multi-sensor fusing approach to self-localization and obstacle perception for robotic cars. *Front. Inf. Technol. Electron. Eng.* **2017**, *18*, 122–138. [CrossRef]
10. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3d object detection network for autonomous driving. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1907–1915.
11. Xu, K.; Yang, Z.; Xu, Y.; Feng, L. A Novel Interactive Fusion Method with Images and Point Clouds for 3D Object Detection. *Appl. Sci.* **2019**, *9*, 1065. [CrossRef]
12. Oh, S.I.; Kang, H.B. Object detection and classification by decision-level fusion for intelligent vehicle systems. *Sensors* **2017**, *17*, 207. [CrossRef] [PubMed]
13. Rakesh, N.R.; Ohn-Bar, E.; Trivedi, M.M. Looking at pedestrians at different scales: A multiresolution approach and evaluations. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 3565–3576.
14. Asvadi, A.; Garrote, L.; Premebida, C.; Peixoto, P.; Nunes, U.J. Multimodal vehicle detection: Fusing 3D-LIDAR and color camera data. *Pattern Recognit. Lett.* **2018**, *115*, 20–29. [CrossRef]
15. Geiger, A.; Lenz, P. 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.
16. Lowe, D.B. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
17. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
18. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
19. Girshick, R. Fast R-CNN. In Proceedings of the IEEE international conference on computer vision(ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
20. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]

21. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
22. Wang, Z.; Zhan, W.; Tomizuka, M. Fusing Bird's Eye View LIDAR Point Cloud and Front View Camera Image for 3D Object Detection. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 1–6.
23. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
24. Paris, S.; Kornprobst, P.; Tumblin, J.; Durand, F. Bilateral Filtering: Theory and Applications. *Found. Trends®Comput. Graph. Vis.* **2008**, *4*, 1–73. [CrossRef]
25. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767v1.
26. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556v6.
27. Janai, J.; Güney, F.; Behl, A.; Geiger, A. Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. *arXiv* **2017**, arXiv:1704.05519.
28. Yebes, J.J.; Bergasa, L.M.; García-Garrido, M.Á. Visual object recognition with 3D-aware features in KITTI urban scenes. *Sensors* **2015**, *15*, 9228–9250. [CrossRef] [PubMed]